Eötvös Loránd University

Faculty of Informatics

Department of Artificial Intelligence

Numerosity-Based Categorization in Neural Networks

Supervisor:
Dr. László Csaba Gulyás
Associate Professor

Author:
Karoki Evans Njogu
Computer Science MSc

Budapest, 2025

Contents

Acknowledgements

I am deeply grateful to my supervisor, Dr. László Csaba Gulyás, for his tremendous guidance throughout this research journey. His sharp insights, tireless attention to detail and constant encouragement not only elevated the quality of this thesis but challenged me to think deeper and reason better. Every time he poked holes in my assumptions, he also handed me the tools to build something stronger and for that, I am sincerely thankful.

I would also like to thank the Department of Artificial Intelligence at Eötvös Loránd University for providing a supportive academic environment and my peers and friends who shared this journey with me, your conversations and presence made the process lighter.

Most importantly, I wish to acknowledge the memory of my late mother, who passed away just ten months ago. Her unwavering faith in me, her love and her prayers were the foundation on which I built not only this thesis, but so much of who I am. This accomplishment is bittersweet without her here, but I carry her spirit with me in every page and every step forward.

To all who walked with me, thank you.

Abstract

Understanding how neural networks can learn to categorize quantities in abstract terms such as *few*, *medium* and *many*, offers a promising direction for bridging cognitive-inspired modeling with modern machine learning. This thesis investigates numerosity-based categorization, aiming to train models that can perceive and generalize numerical groupings without relying on explicit counting or object identity.

We explore this through a multi-phase experimental design. First, convolutional neural networks (CNNs) are trained on synthetic dot-pattern images representing varying quantities. We then test the models' ability to generalize to silhouette-based images composed of unrelated real world object classes such as apples, various animals and common everyday objects. Finally, to assess whether models are learning genuine numerosity or merely reacting to surface-level visual cues, we construct and evaluate a pixel-ratio-controlled dataset with structured and random white-pixel distributions.

Results demonstrate strong within-domain classification performance, with CNN and CNN+Transformer hybrids showing promising generalization across modalities. However, the experiments also reveal limitations in generalizing when low-level visual cues are tightly controlled. These findings highlight the challenges of abstract quantity recognition and open avenues for exploring cognitively plausible architectures that better mimic human perception of numerosity.

This work contributes to the understanding of abstract categorization in neural networks, particularly in contexts where exact counting is impractical. Applications include crowd estimation, traffic monitoring and educational tools; domains where the question is often not "how many exactly," but rather "approximately how many."

Chapter 1

Introduction

Humans have the natural ability to estimate the number of objects in a scene or environment without consciously counting. This intuitive skill, known as numerosity perception, enables us to quickly distinguish between a few, several or many objects in our environment. Unlike explicit counting, numerosity perception is approximate, abstract and often sufficient for decision-making in daily life. Replicating this capability in artificial neural networks presents a unique challenge: how can machines learn to recognize and generalize quantities without relying on exact enumeration or prior knowledge of object categories?

This thesis explores numerosity-based categorization in neural networks, specifically the ability to group inputs into abstract categories such as *few*, *medium* and *many*. The goal is to understand whether neural models can develop flexible, perceptual representations of quantity that extend across different visual domains.

The research begins with synthetic dot-pattern images, which offer a controlled baseline for training convolutional neural networks (CNNs) on numerical abstraction. From there, the study investigates whether models trained on dot patterns can generalize to images constructed from silhouettes of real-world items, such as apples, various animals and common objects like hammers, shoes, pens etc. These silhouette-based datasets introduce semantic and visual variability while keeping the numerosity signal intact.

Beyond generalization, the thesis conducts a series of transfer learning experiments. These include fine-tuning dot-trained models on silhouette datasets, training new models from scratch on the silhouette data and performing the reverse procedure to examine the transferability of learned numerosity features in both directions. This

broader evaluation enables a more comprehensive analysis of model adaptability and representation learning.

To probe whether models are genuinely learning abstract numerosity or merely responding to surface-level features such as pixel density, a final experiment introduces pixel-controlled datasets. These images maintain fixed white pixel ratios across structured and randomized patterns, decoupling quantity from obvious visual cues.

Together, these experiments form a multi-layered investigation into how neural networks perceive, represent and generalize abstract quantity. The findings contribute to our understanding of cognitive-inspired modeling in artificial systems and highlight the potential and current limitations of neural networks in learning human-like numerosity intuition.

1.1 Problem Statement

Despite the impressive capabilities of deep learning in tasks such as object detection and image classification, neural networks typically rely on explicit labeling and category memorization. When it comes to quantity perception, models are often trained for exact counting, which is both data-intensive and unnecessarily precise for many real-world applications. However, in daily life, humans often reason about quantity in abstract terms (identifying groups as “few,” “some,” or “many”) without needing to know the exact number.

This thesis addresses the gap between such cognitive-inspired abstraction and current machine learning approaches by exploring whether neural models can learn to perceive numerosity independently of object identity and visual complexity. The central problem lies in determining if abstract quantity representations can emerge in neural networks trained on perceptual input, and whether those representations can generalize across different visual domains.

Furthermore, the work investigates whether models are truly learning numerical abstraction or are instead relying on low-level proxies, such as pixel density or area coverage. By designing a variety of datasets and transfer-learning experiments, this thesis examines the degree to which models internalize numerosity as a meaningful, generalizable feature.

1.2 Motivation

Humans have a great ability to estimate quantities quickly and approximately, an ability that is believed to be rooted in evolutionary survival needs such as detecting groups of predators or food items. This perceptual skill often referred to as subitizing or numerosity estimation is fundamentally different from precise counting. It allows for fast, rough judgments that are often sufficient and more efficient in real-world scenarios.

In contrast, artificial neural networks typically approach quantity perception as a regression or classification task tied to exact numbers. While this may be suitable for some applications, many practical domains such as crowd monitoring, ecological surveys and early educational tools do not require precise counts. Instead, the ability to categorize numerical information into abstract groupings like "few," "medium," or "many" is not only more flexible but also more cognitively plausible.

This thesis is motivated by the question of whether such abstraction can emerge in neural networks. By focusing on numerosity as a perceptual concept rather than a strict numerical value, we aim to bridge the gap between machine learning and human-like reasoning. The ability of a model to generalize across visually diverse domains, while maintaining an abstract sense of quantity, has implications for both cognitive modeling and real-world AI systems that operate under uncertainty or limited supervision.

1.3 Research Objectives

The main objective of this thesis is to investigate whether neural networks can develop an abstract sense of quantity or numerosity through perceptual input, without relying on explicit counting or object recognition.

To support this aim, the thesis pursues the following specific objectives:

- To train convolutional neural networks (CNNs) to categorize synthetic dot-pattern images into abstract numerical classes such as *few*, *medium* and *many*.
- To evaluate the ability of dot-trained models to generalize to silhouette-based images composed of real-world items such as apples, animals and tools.

- To fine-tune dot-trained models using the silhouette dataset and assess the improvement in cross-domain performance.
- To train new models from scratch on the silhouette dataset and compare their performance with fine-tuned models.
- To perform reverse experiments by fine-tuning silhouette-trained models on dot patterns, thereby examining bidirectional transferability.
- To investigate whether neural networks are learning true numerosity or relying on low-level features by introducing a pixel-ratio-controlled dataset.
- To analyze and compare model performance across all settings using classification metrics and visualizations such as confusion matrices and accuracy trends.

1.4 Thesis Structure

This thesis is organized into seven chapters, each of them addressing a specific aspect of the research:

- **Chapter 1 – Introduction:** Provides the background, motivation, problem statement and objectives of the research.
- **Chapter 2 – Related Work:** Reviews existing literature on numerosity perception, concept learning in neural networks and approaches to generalization and abstraction in machine learning.
- **Chapter 3 – Proposed Solution:** Describes the conceptual approach and design decisions made for numerosity-based categorization, including dataset construction and abstraction targets.
- **Chapter 4 – Methods:** Details the experimental setup, including datasets (dot patterns, silhouette images and pixel-ratio variants), neural architectures used, training procedures and evaluation metrics.
- **Chapter 5 – Results:** Presents the outcomes of all experiments, including generalization tests, fine-tuning results, reverse transfer experiments and performance on controlled pixel datasets.

- **Chapter 6 – Discussion:** Interprets the results in the context of cognitive plausibility, model limitations and the role of visual features versus abstract representations.
- **Chapter 7 – Conclusion and Future Work:** Summarizes the key findings of the thesis and suggests potential directions for future research on cognitive-inspired learning in neural networks.

Chapter 2

Related Work

2.1 Cognitive Perspectives on Numerosity

Numerosity perception refers to the human ability to rapidly and approximately gauge the number of objects in a visual scene without explicit counting. This capacity, often termed subitizing, is observable even in infancy and across various animal species [2, 12]. Cognitive studies have shown that this skill is distinct from symbolic arithmetic, relies on approximate rather than exact quantities, and is processed in dedicated neural regions such as the intraparietal sulcus [13, 15].

The approximate number system (ANS) underlies this capability, enabling rough quantity estimation that is robust to changes in object size, shape and spacing. Subitizing, typically effective for small quantities (≤ 4), is fast, accurate and evolutionarily advantageous. Neuroimaging studies have identified a distributed network involving the intraparietal sulcus (IPS), dorsolateral prefrontal cortex (DLPFC) and visual cortices as key regions activated during non-symbolic number processing [7]. Individual differences in numerosity acuity have been linked to both gray matter volume and functional activation patterns in these areas [22].

2.2 Numerosity in Neural Networks

Early computational models attempted to emulate numerosity through hand-engineered mechanisms or recurrent structures inspired by cognitive theories. For example, models using on-center off-surround dynamics or progressive unsupervised

learning frameworks [19] attempted to simulate developmental refinement of numerical acuity.

Subsequent studies demonstrated that deep convolutional neural networks (CNNs), even when trained for unrelated visual tasks, can spontaneously develop number-selective units [6]. However, these networks often relied heavily on superficial features such as total area or object density, limiting their robustness and generalization [11]. Efforts to mitigate these deficits include training with synthetic datasets designed to decorrelate visual cues from numerical content and leveraging recurrent structures to model sequential counting or inhibition-based tuning curves.

This limitation was further examined by Wu et al. [21], who conducted a series of cognitive-style experiments to probe whether deep learning can truly internalize abstract number concepts like humans. Their findings revealed that standard deep networks fail to generalize the notion of numerosity when tested on visually varied representations, suggesting a fundamental cognitive deficit in black-box learning. The study also introduced a recurrent convolutional model with morphological priors, capable of subitizing deterministically, underscoring the value of embedding cognitive structure into deep models for abstract quantity tasks.

2.3 Transformer Models and Global Reasoning in Vision

Transformer-based models, originally introduced for natural language processing, have been successfully extended to vision tasks through the Vision Transformer (ViT) [4]. Unlike CNNs that operate on local receptive fields, for transformers to capture global relationships between image regions, they rely on self-attention.

This property makes transformers especially promising for numerosity tasks: abstracting "quantity" requires recognizing object groupings across varied spatial layouts, not merely local texture [17].

2.4 Transformers for Numerosity Categorization

Recent studies demonstrate that transformers outperform CNNs under distributional shifts, due to their global receptive fields and more uniform internal represen-

tations [10]. This allows them to better generalize abstract concepts like numerosity across domains.

While explicit applications of transformers to numerosity categorization are still rare, research indicates that attention-based architectures can better capture abstract structure, supporting flexible quantity reasoning [20].

2.5 Approximate Counting and Developmental Learning

Stoianov and Zorzi [19] showed that unsupervised learning could lead to spontaneous number-selective units, simulating aspects of developmental cognitive trajectories. More recently, progressive stochastic learning frameworks have been proposed to model how discrimination acuity improves over time [15], echoing human cognitive development.

Recent research has also explored integrating symbolic priors into neural networks to support more human-like subitizing. Alam et al. [1] introduced a neuro-symbolic loss function using Holographic Reduced Representations (HRR) to enhance generalization in numerosity tasks. Their model demonstrated improved robustness to structural variation, offering a promising direction for combining perception with structured symbolic reasoning.

Fine-tuning and domain transfer techniques have also been explored to promote numerosity abstraction across varied inputs.

2.6 Density Estimation and Fuzzy Quantifiers

Other studies frame numerosity classification as a type of density estimation. Techniques based on generative modeling and density maps [16, 14] offer frameworks for learning spatial layouts without explicit counting.

In parallel, fuzzy quantifiers provide a soft logic approach to categorizing "few," "medium," and "many," enabling neural models to handle noisy and overlapping categories [8].

2.7 Crowd Counting and Object Density Estimation

Crowd counting is a real-world parallel to numerosity research. Density map prediction methods such as Hydra CNN and CCNN have been used to estimate people counts under occlusion and perspective shifts [18, 3]. Although these models aim at exact counting, their handling of dense spatial distributions informs the design of models for approximate numerosity.

2.8 Concept Learning and Meta-Learning

Recent advances in concept learning and meta-learning [5] offer principles for training models to generalize abstract concepts beyond specific datasets.

Symbolic integration methods, such as HRR loss in subitizing tasks [1], seek to improve neural representations of structure, offering another pathway toward flexible quantity understanding.

2.9 Positioning of This Thesis

While previous research has addressed either cognitive modeling of numerosity or the technical implementation of counting in neural systems, relatively few have tackled the problem of categorizing visual inputs into abstract numerical classes across diverse domains. This thesis contributes by:

- Investigating whether neural networks can generalize the concepts of "few," "medium," and "many" across dot patterns, object silhouettes and pixel-based visual abstractions.
- Exploring how effective transfer learning and fine-tuning is between these modalities.
- Evaluating whether classification is driven by abstract numerosity or surface-level cues like pixel ratios.
- Comparing convolutional and transformer-based models for their ability to generalize numerical abstraction across visually distinct domains.

- Providing insights into the strengths and limitations of neural architectures in modeling quantity abstraction, guided by principles of cognitive plausibility.

Chapter 3

Proposed Solution

This thesis proposes a numerosity-based classification framework that categorizes visual inputs into abstract quantity classes (*few*, *medium* and *many*) across different visual domains. Unlike traditional counting models that focus on precise enumeration, the proposed approach seeks to reflect how humans estimate quantity in real-world scenarios: approximately, flexibly and invariant to object identity.

The central hypothesis is that deep neural models can learn to abstract numerosity from perceptual input alone, without relying on object-specific features. By training models to map images of varying item counts to coarse numerical categories, this work aims to replicate a form of visual quantification akin to the approximate number system (ANS) observed in humans.

The proposed solution is built on the following design principles:

- **Abstraction over Counting:** The goal is to learn broad quantity classes rather than exact numerical counts, mirroring the approximate nature of human numerosity perception.
- **Modality-Agnostic Generalization:** The framework evaluates whether models can generalize quantity abstraction across visually distinct domains: dot patterns, silhouettes of real-world objects (e.g., apples, animals, hammers) and synthetic pixel-density patterns.
- **Architectural Diversity:** Both convolutional neural networks (CNNs) and hybrid CNN + Transformer models are explored to assess how different architectural biases affect abstraction and generalization.

- **Transfer Learning and Fine-Tuning:** Experiments probe whether a model trained on one modality (e.g., dots) can be successfully fine-tuned or evaluated on another (e.g., silhouettes), assessing representational flexibility.
- **Cognitive Plausibility:** Experimental setups reflect cognitive theories of numerosity, testing model robustness to variations and evaluating reliance on meaningful quantity abstraction rather than superficial surface cues.

The implementation follows three core components:

1. **Dataset Pipelines:** Custom dataset generators produce labeled samples for dot patterns, object silhouettes and pixel-ratio controlled images. Each dataset is split into training, validation and test sets, with predefined thresholds for categorizing into *few*, *medium* and *many*.
2. **Model Architectures:** Two model types are implemented: (1) a lightweight CNN and (2) a hybrid CNN + Transformer architecture. Both are evaluated across datasets for in-domain accuracy and cross-domain generalization.
3. **Experiment Logic:** Experiments are structured to test within-modality learning, cross-modality transfer and abstraction quality. Special emphasis is placed on evaluating whether models learn true numerosity or rely on proxies like pixel density.

This chapter outlines the rationale behind the proposed solution. The following chapter details the methods used, including dataset generation, architecture configurations, training protocols and evaluation strategies.

Chapter 4

Methods

4.1 Overview

This chapter outlines the experimental design used to evaluate numerosity-based categorization in artificial neural networks. The central objective was to train a model capable of classifying visual input into abstract quantity categories: *few*, *medium* and *many* and to assess its ability to generalize under both controlled variations and cross-domain shifts.

Experiments were conducted in a staged manner:

- First, a baseline model was trained on synthetic grayscale dot pattern images.
- Next, the model was tested on variants of the dot data to assess within-modality generalization under shape, occlusion and spatial clustering perturbations.
- Cross-modality generalization was investigated through both evaluation and training on silhouette-based images, including transfer learning in both directions between dot and silhouette modalities.

Both convolutional and hybrid architectures were considered. Evaluation focused on classification accuracy, generalization drop and qualitative analysis using embedding visualizations.

4.2 Dataset Generation

4.2.1 Dot Pattern Dataset (Baseline)

The dot pattern dataset was synthetically generated to serve as a clean and controlled baseline for learning abstract numerosity categories. Each image is a grayscale square of size 128×128 pixels, containing randomly placed white dots on a black background. The dataset comprises 5000 samples, categorized into three classes based on dot count:

- **Few:** 1–5 dots
- **Medium:** 6–15 dots
- **Many:** 16–30 dots

Generation Process. Each image was initialized as a blank matrix. A random number of dots (between 1 and 30) was rendered on the image. For each dot:

- The center (x, y) coordinates were sampled such that the full dot stayed within the image bounds.
- The radius was randomly chosen between 3 and 12 pixels.
- Dots were drawn using filled white circles (grayscale value 255) via OpenCV.

To promote robustness during training, the dataset was designed to support optional augmentations such as Gaussian noise, brightness variation and mild rotations, though these were not applied during the main experiments.

Preprocessing. All images were normalized to the $[0, 1]$ range and stored as PyTorch tensors with shape $(1, 128, 128)$. Labels were encoded as integers: 0 for "Few", 1 for "Medium" and 2 for "Many", following the convention used in the PyTorch training pipeline.

Splitting and Storage. The dataset was split into 70% training, 15% validation and 15% test sets. All subsets were serialized using `torch.save()` to ensure reproducibility and consistency across experimental runs.

Representative examples from each class in the Dot Pattern Dataset are shown in Figure 4.1.

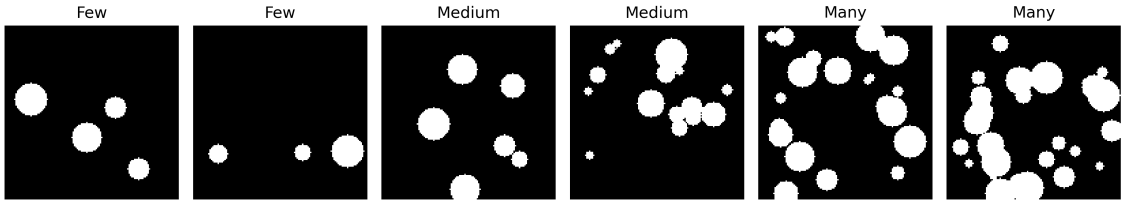

Figure 4.1: Examples from the Dot Pattern Dataset showing the three numerosity categories: Few (1–5), Medium (6–15) and Many (16–30). Each image is 128×128 pixels with randomized dot sizes and positions.

Justification. Dots were chosen based on their wide usage in the numerosity literature and their simplicity. Unlike real-world object datasets, dot patterns reduce confounding factors and enable focused exploration of numerosity perception. The use of grayscale further simplifies input and computational load, while preserving the essential spatial features required for learning abstract quantity categories.

4.2.2 Within-Modality Variant Datasets

To assess whether the model learned abstract numerosity or relied on superficial visual patterns, several controlled variants of the Dot Pattern Dataset were generated. These datasets share the same format and label distributions but introduce perceptual challenges that simulate real-world variations.

Shape Variant. Dots were replaced with alternative geometric primitives such as squares and triangles. Shapes were rendered using OpenCV polygonal approximations with randomly sampled sizes and orientations. The goal was to test if the network associates numerosity with overall shape or with discrete item count.

Occlusion Variant. This variant introduced deliberate overlaps between shapes. Each object was assigned a probability of overlapping with another. By introducing ambiguity in object boundaries, this setup tested whether the model could still infer correct quantity in visually ambiguous scenes.

Clustered Variant. Instead of uniform random placement, objects were arranged in tight spatial clusters or irregular groupings. A small number of cluster centers were

randomly chosen and multiple items were sampled around them using Gaussian jitter. This simulates the perceptual challenge of segmenting objects in crowded or non-uniform layouts.

Visual Samples. Examples from the Shape, Occlusion and Clustered variants are presented in Figure 4.2, illustrating the types of distortions applied while keeping numerosity labels consistent.

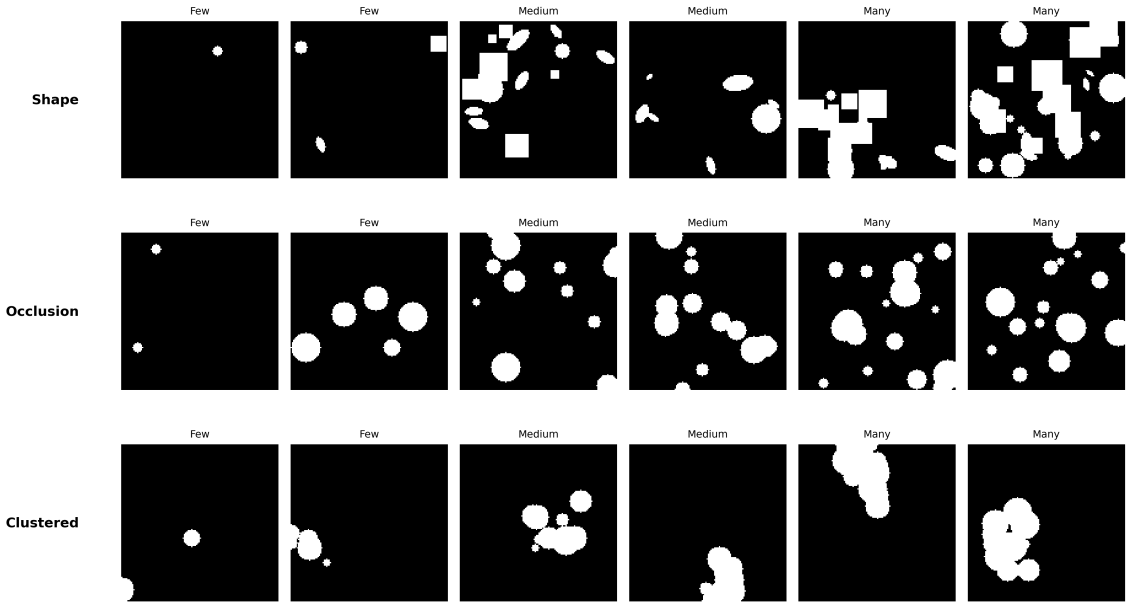

Figure 4.2: Dot Pattern Dataset Variants with class-wise samples. Each row shows images from a distinct variant (Shape, Occlusion, Clustered), with two samples per numerosity class (Few, Medium, Many). The goal is to test the model's cability to generalize abstract quantity under visual distortions.

Label Consistency. In all three variants, the number of items per image remained unchanged from the original dot dataset (i.e., "Few", "Medium" and "Many" based on count thresholds of 1–5, 6–15 and 16–30 respectively). This ensured that only the visual structure varied while the semantic target (numerosity class) stayed constant.

Purpose. These variations serve as controlled tests for within-modality generalization, aiming to reveal whether a trained model relies on item count alone or becomes sensitive to irrelevant features such as shape identity, spatial uniformity, or occlusion cues.

4.2.3 Silhouette-Based Numerosity Dataset (MPEG-7 Inspired)

To evaluate the ability of the model to generalize numerosity perception across semantically richer and perceptually diverse inputs, a new dataset was generated using silhouettes from the MPEG-7 CE-Shape-1 dataset. This dataset includes binary representations of real-world categories such as animals, tools and shapes (e.g., *apple, elephant, hammer, butterfly*), serving as a more realistic and cognitively demanding alternative to dot patterns.

About MPEG-7 CE-Shape-1. The MPEG-7 CE-Shape-1 dataset, curated by Dr. Longin Jan Latecki, was originally developed to benchmark algorithms for shape retrieval and matching. It contains 70 object categories, each represented by 20 high-contrast binary silhouette images. The dataset was carefully designed to introduce intra-class variability through rotation, articulation and partial occlusion, providing a rich source of perceptual diversity for shape-based recognition tasks [9]. Figure 4.3 illustrates representative examples from the original collection.

Dataset Examples. Figure 4.3 shows example object classes from the MPEG-7 CE-Shape-1 Test Set, illustrating the diversity of shapes used for constructing the silhouette-based numerosity dataset.

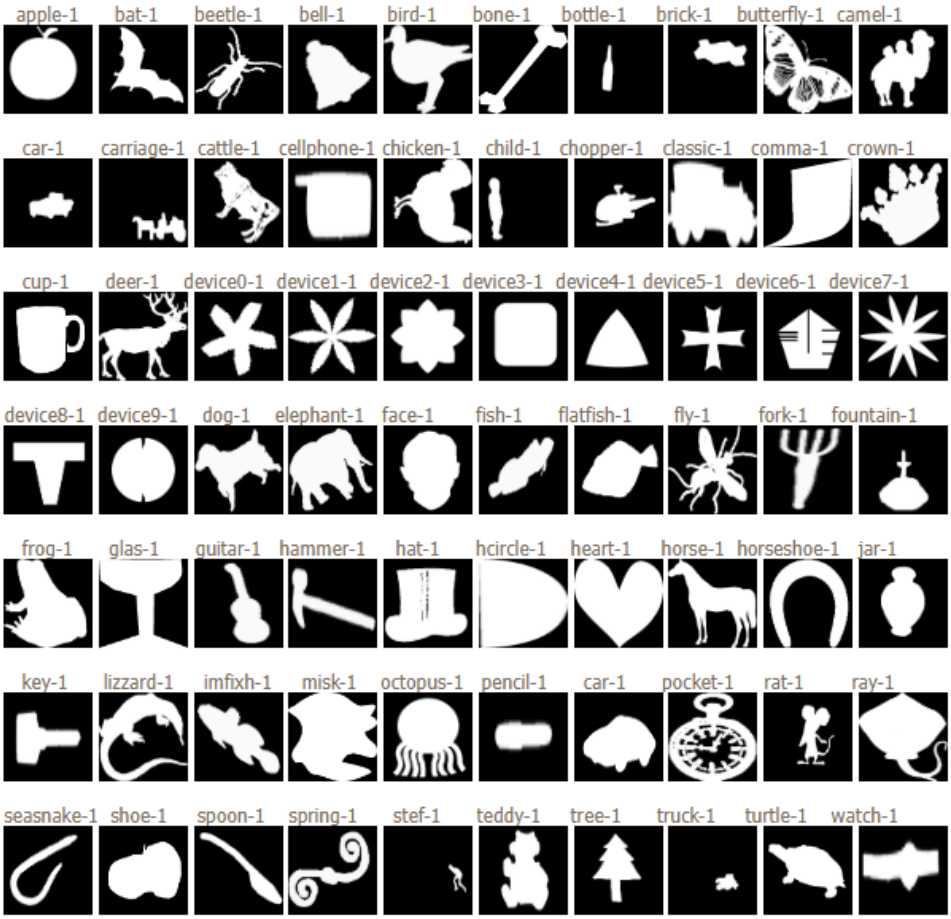

Figure 4.3: Example classes from the MPEG-7 CE-Shape-1 Test Set used in constructing the silhouette-based numerosity dataset. The dataset includes a wide range of semantically meaningful and perceptually diverse object shapes.

Preprocessing and Source Handling. The original binary GIF files were processed to extract only the first frame. Images were converted to grayscale, resized to fit within 40×40 pixels while preserving aspect ratio and binarized to ensure high contrast.

Composite Image Generation. Each image in the dataset was synthesized by randomly compositing silhouette objects onto a blank 128×128 canvas. The number of silhouettes placed per image was randomly sampled between 1 and 30. The process followed these steps:

- Select a silhouette shape from the pool.
- Resize it randomly (typically between 20×20 and 40×40).
- Randomly sample a non-overlapping location on the canvas.

- Paste the silhouette using OpenCV masking operations.

This process ensured that the number of objects per image governed the label, not their identity, size, or class.

Labeling and Categorization. Numerosity was mapped to three abstract classes using consistent thresholds:

- **Few:** 1–5 objects
- **Medium:** 6–15 objects
- **Many:** 16–30 objects

This categorization mirrored the dot dataset to maintain label consistency across modalities.

Dataset Structure and Storage. A total of 3000 images were generated and split into 70% training, 15% validation and 15% test sets. The images were stored as PNG files and the corresponding PyTorch datasets were serialized using `torch.save()` to ensure reproducibility across experiments.

Dataset Samples. Representative samples from the silhouette-based numerosity dataset are shown in Figure 4.4, illustrating the variation in object placement and numerosity across the Few, Medium and Many categories.

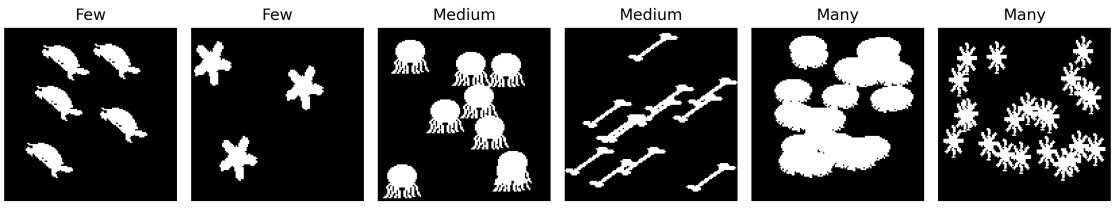

Figure 4.4: Sample images from the silhouette-based numerosity dataset. Each image corresponds to an abstract quantity class based on the number of silhouette objects: Few (1–5), Medium (6–15) and Many (16–30). Shapes are sampled from real-world categories in the MPEG-7 dataset.

Justification and Challenges. The choice of silhouettes serves two key purposes:

1. **Semantic richness:** Shapes are meaningful and vary in contour, topology and form, introducing real-world variation without texture or color.

2. **Abstract generalization:** The model must ignore identity-specific cues and focus on global numerosity estimation.

Unlike dots, silhouette objects vary in complexity and can partially occlude one another. This makes the dataset more cognitively and computationally challenging, simulating real-world conditions under which humans perform approximate quantity estimation. Successful generalization on this dataset would suggest that the model has internalized a truly abstract, identity-invariant concept of numerosity.

4.2.4 Pixel-Ratio Controlled Dataset

To investigate whether the neural model learns abstract numerosity or instead exploits low-level image statistics such as pixel density, a synthetic dataset was constructed where images encode quantity purely through the ratio of white pixels to total image area. This dataset eliminates discrete objects and introduces structured patterns of intensity, allowing us to isolate and evaluate the model's sensitivity to raw visual mass.

Motivation. This dataset serves as a diagnostic tool to test whether the model's numerosity predictions are grounded in perceptual abstraction or driven by simple heuristics such as total brightness or area. Unlike the dot and silhouette datasets, these images do not contain recognizable object boundaries.

Generation Strategy. Each image is 128×128 pixels in size and belongs to one of three classes, defined by target white pixel ratios:

- **Few:** 20% white pixels
- **Medium:** 50% white pixels
- **Many:** 80% white pixels

The white pixels were arranged using four distinct spatial patterns:

1. **Vertical Bands:** White columns stacked on the right
2. **Horizontal Bands:** White rows stacked on the bottom
3. **Checkerbox:** Structured grid of 8×8 white blocks
4. **Random Noise:** White pixels scattered randomly across the canvas

These patterns allow us to control both the total pixel ratio and spatial layout, probing whether spatial structure influences prediction in the absence of countable objects.

Labeling and Consistency. Each image was labeled according to its white pixel ratio using thresholds consistent with earlier datasets:

- Label 0: **Few** (0.2 ratio)
- Label 1: **Medium** (0.5 ratio)
- Label 2: **Many** (0.8 ratio)

This ensured that models trained on object-based datasets could be evaluated on this abstraction-free dataset without relabeling.

Dataset Composition and Storage. The dataset includes 1,200 total samples:

- 100 images per class for each of the 4 patterns
- Split 70% training, 15% validation, 15% test

All images were stored as grayscale tensors normalized to $[0, 1]$ and serialized as PyTorch objects using `torch.save()`.

Dataset Samples. Figure 4.5 illustrates examples from the Pixel-Ratio Controlled Dataset, highlighting the structured spatial patterns (vertical bands, horizontal bands, checkerbox and random scatter) across the abstract numerosity categories.

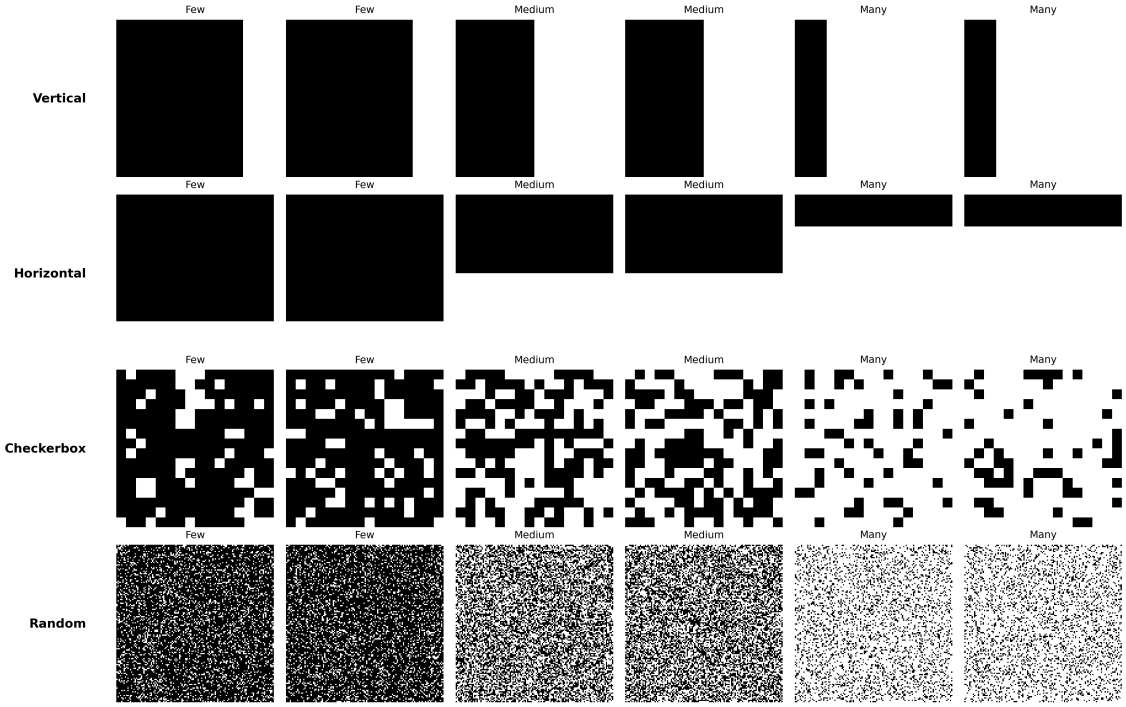

Figure 4.5: Samples from the Pixel-Ratio Controlled Dataset. Each row corresponds to a spatial pattern type (Vertical, Horizontal, Checkerbox, Random). Columns show two images per numerosity class: Few (20% white pixels), Medium (50%) and Many (80%). These synthetic abstractions allow evaluation of models without reliance on object-based cues.

Justification. By removing object identity entirely and relying only on the distribution of white pixels, this dataset serves as a control condition for probing whether a model's concept of numerosity is based on quantity itself or surrogate statistics like brightness or area. Performance drops on this dataset would suggest reliance on object-based reasoning, while stable accuracy would imply a more general understanding of quantity grounded in spatial abstraction.

4.3 Model Architecture

Two neural network architectures were developed and evaluated in this study: a baseline convolutional model and a hybrid CNN + Transformer model. Both architectures share a common goal, to categorize input images into abstract numerosity classes: *Few*, *Medium* and *Many*. Their designs differ in how they capture local versus global visual dependencies.

4.3.1 Residual CNN Architecture

The baseline model, referred to as `NumerosityCNN`, is a convolutional neural network built using residual blocks. It consists of three convolutional stages with increasing depth and resolution reduction, followed by a fully connected head for classification.

Architecture.

- **Input:** $1 \times 128 \times 128$ grayscale image
- **Layer 1:** ResidualBlock($1 \to 32$) + MaxPool(2×2)
- **Layer 2:** ResidualBlock($32 \to 64$) + MaxPool(2×2)
- **Layer 3:** ResidualBlock($64 \to 128$) + MaxPool(2×2)
- **Flatten:** Output feature map of size $128 \times 16 \times 16$ flattened
- **FC1:** Fully connected layer ($32768 \to 128$), ReLU
- **FC2:** Linear layer ($128 \to 3$), representing numerosity classes

Residual Block Design. Each residual block includes two convolutional layers with batch normalization and a shortcut connection. If the input and output dimensions differ, a 1×1 convolution is used in the shortcut.

Purpose. This architecture serves as a fast and interpretable baseline. It focuses on capturing local features and spatial clustering patterns.

Architecture Overview. Figure 4.6 illustrates the structure of the baseline NumerosityCNN model, which combines three residual convolutional blocks with pooling layers, followed by fully connected layers for classification into abstract numerosity classes.

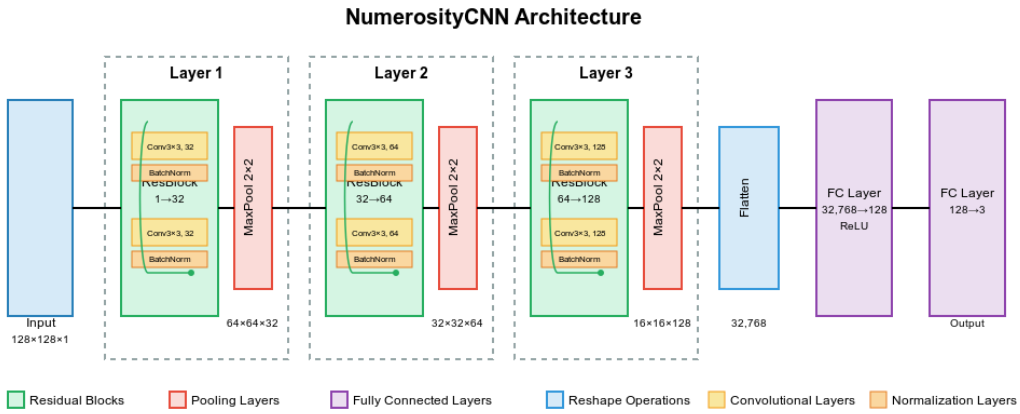

Figure 4.6: Architecture of the NumerosityCNN model. The network consists of three residual convolutional blocks followed by max-pooling and two fully connected layers. The design emphasizes local spatial learning and efficient parameterization.

4.3.2 CNN + Transformer Hybrid Architecture

To introduce global contextual reasoning, a hybrid model named `NumerosityCNNTransformer` was developed. It combines a CNN backbone with a lightweight Vision Transformer head.

CNN Backbone. The first part of the model mirrors the baseline CNN up to Layer 3, producing a $128 \times 16 \times 16$ feature map.

Patch Embedding.

- The feature map is patchified using a 4×4 convolution (stride 4), producing $8 \times 8 = 64$ patches.
- Each patch is embedded into a 128-dimensional vector.
- Positional encodings are added to preserve spatial information.

Transformer Encoder.

- Two encoder layers are used with 4 attention heads.
- Each layer includes multi-head self-attention and a feed-forward subnetwork with residual connections.
- The final token sequence is mean-pooled.

Classification Head.

- The pooled token output is layer-normalized and passed through a dropout layer ($p = 0.4$).
- A final linear layer maps it to 3 class logits.

Motivation. This hybrid model allows us to test whether spatial self-attention improves numerosity classification in scenarios requiring global reasoning such as with silhouettes or pixel-ratio patterns.

Hybrid Model Overview. Figure 4.7 presents the NumerosityCNNTransformer architecture. It combines a CNN backbone for local feature extraction with a Transformer encoder to enable global spatial reasoning, before final classification into numerosity categories.

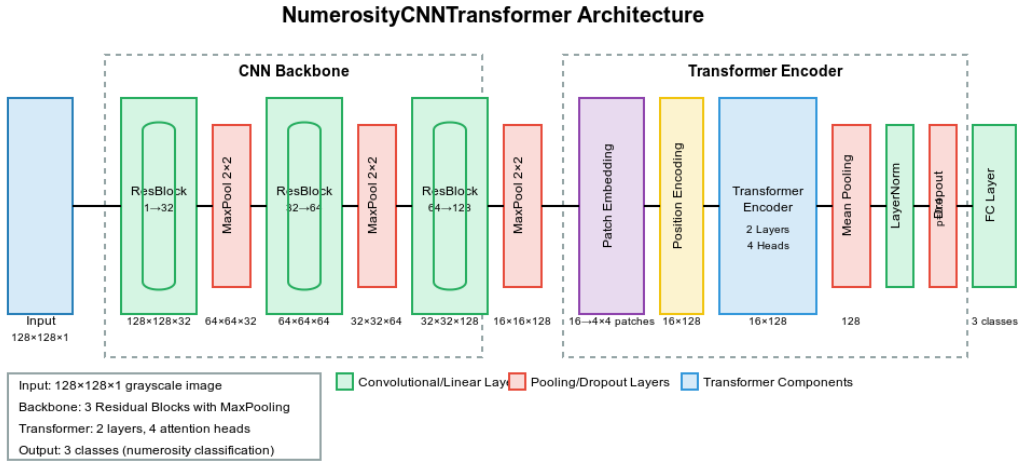

Figure 4.7: Architecture of the hybrid NumerosityCNNTransformer model. A CNN backbone extracts spatial features which are patchified and passed to a Transformer encoder with positional encoding and self-attention layers. The output is pooled and classified into numerosity categories.

4.4 Training Configuration and Hyperparameter Selection

All models were trained using the PyTorch framework. The following choices were made regarding the training setup and hyperparameters:

Optimizer. The AdamW optimizer was selected for all experiments after preliminary trials indicated improved stability and generalization compared to vanilla Adam. AdamW's decoupled weight decay has been shown to better regularize deep models.

Learning Rate. Initial learning rate settings were based on manual tuning. A learning rate of 0.001 was used for early experiments, but tuning revealed that 0.0002 produced better convergence on silhouette datasets, where the input variability is higher. Final settings varied between 0.0001 and 0.0002 depending on the model and dataset.

Batch Size. Batch size was adjusted based on available memory and training stability. Models on the Dot Dataset initially used a batch size of 64, later increased to 128 or 256 for improved gradient estimates and smoother training, particularly on silhouette data.

Dropout. Dropout was introduced in deeper layers to prevent overfitting. Values between 0.3 and 0.4 were manually tuned based on validation set performance.

Weight Decay. A small weight decay (5×10^{-4}) was applied in models trained with AdamW to further promote generalization, especially during experiments involving larger batch sizes.

Early Stopping and Epoch Limits. Training was monitored using validation accuracy and loss curves. Early stopping was implemented manually by reviewing training plots and selecting model checkpoints prior to overfitting onset (usually between 30 and 50 epochs).

Rationale. Hyperparameters were tuned through a combination of:

- Manual search based on validation set performance
- Prior experience with CNN and Transformer architectures
- Computational resource constraints (batch size, training time)

No exhaustive grid search was conducted, as the focus was on robust generalization trends rather than marginal hyperparameter optimization.

Chapter 5

Results

5.0.1 Experiment 1: CNN on Dot Pattern Dataset

This experiment establishes a baseline using a convolutional neural network (CNN) trained on synthetic dot-pattern images. The goal is to test whether a simple CNN can learn to abstractly categorize images into "Few", "Medium" and "Many" classes based on numerosity alone.

Table 5.1 presents the per-class precision, recall and F1-score metrics. The model achieves strong recall and precision across all three categories, especially for the "Many" class, indicating high performance in discriminating high-count inputs. The total support (750) reflects the number of samples evaluated in the test set. The final row shows macro-averaged scores across all classes.

Table 5.1: Final Test Set Accuracy and Per-Class Metrics for Experiment 1 (CNN on Dot Dataset). CNN achieves high recall and precision across all classes, especially for the "Many" category.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Few	0.89	**0.93**	0.91	122
Medium	0.89	0.83	0.86	277
Many	**0.91**	**0.94**	**0.92**	351
Average	**0.90**	**0.90**	**0.90**	**750**

Table 5.2 reports the final overall test accuracy, which stands at 89.73%, consistent with the class-wise metrics.

Table 5.2: Overall Test Set Accuracy for Experiment 1 (CNN on Dot Dataset).

Metric	**Value**
Final Test Accuracy	**89.73%**

Figure 5.1 illustrates the training and validation curves. The left subplot shows that both training and validation loss decrease over time, although some fluctuation is visible near the end, which may suggest minor overfitting. The right subplot confirms stable validation accuracy above 89% after early epochs.

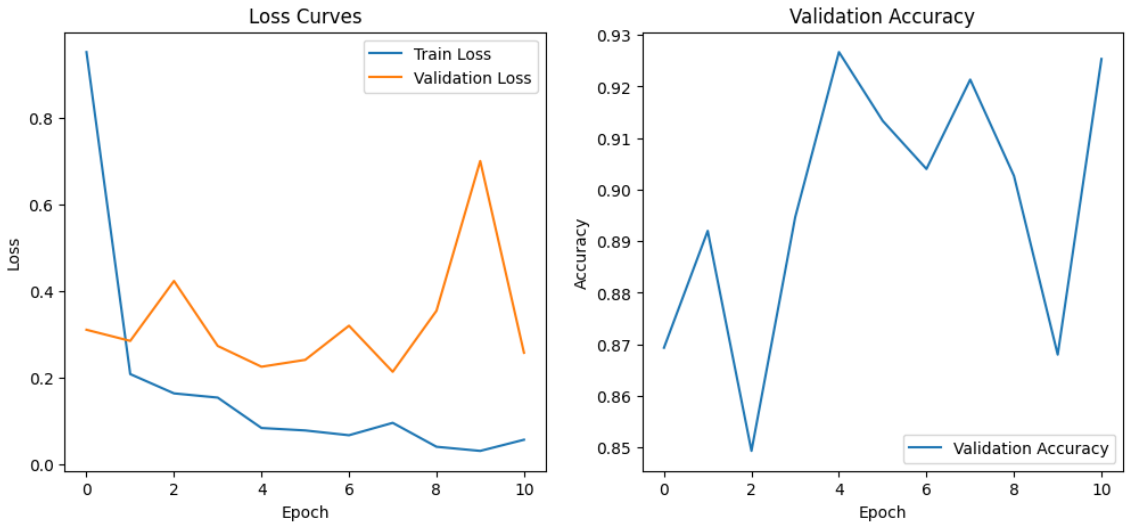

Figure 5.1: Training and validation curves for Experiment 1 (CNN on Dot Dataset). Left: Train vs Validation Loss. Right: Validation Accuracy over epochs.

Figure 5.2 provides the normalized confusion matrix (in percentages). The highest confusion is observed in the "Medium" class, which tends to be misclassified as either "Few" or "Many." In contrast, "Few" and "Many" categories achieve strong separation.

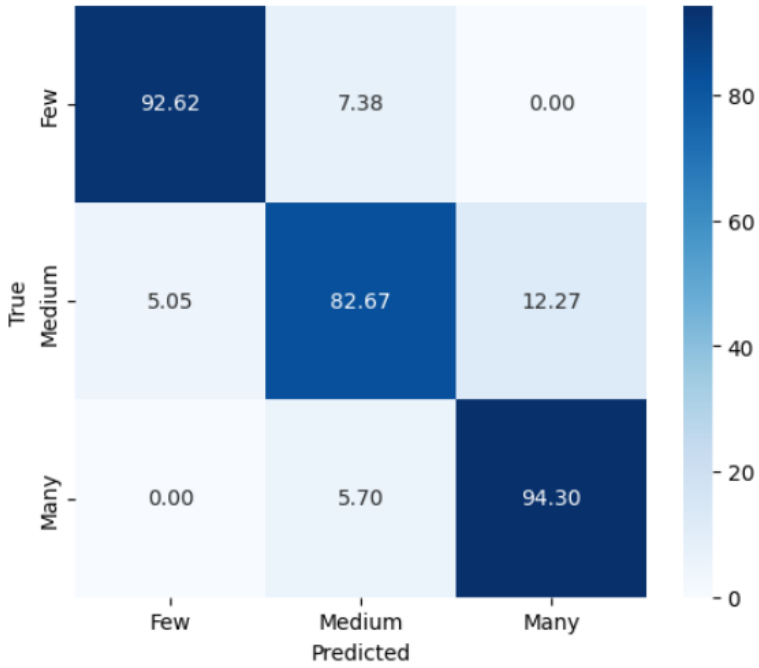

Figure 5.2: Confusion matrix for Experiment 1 (CNN on Dot Dataset). The highest confusion occurs between "Medium" and neighboring categories, but "Few" and "Many" classes are well separated. Values are percentages.

Figure 5.3 shows representative test samples with predicted labels. Green labels indicate correct predictions, while red labels mark misclassifications. Most errors occur near class boundaries, reaffirming the quantitative findings in the confusion matrix.

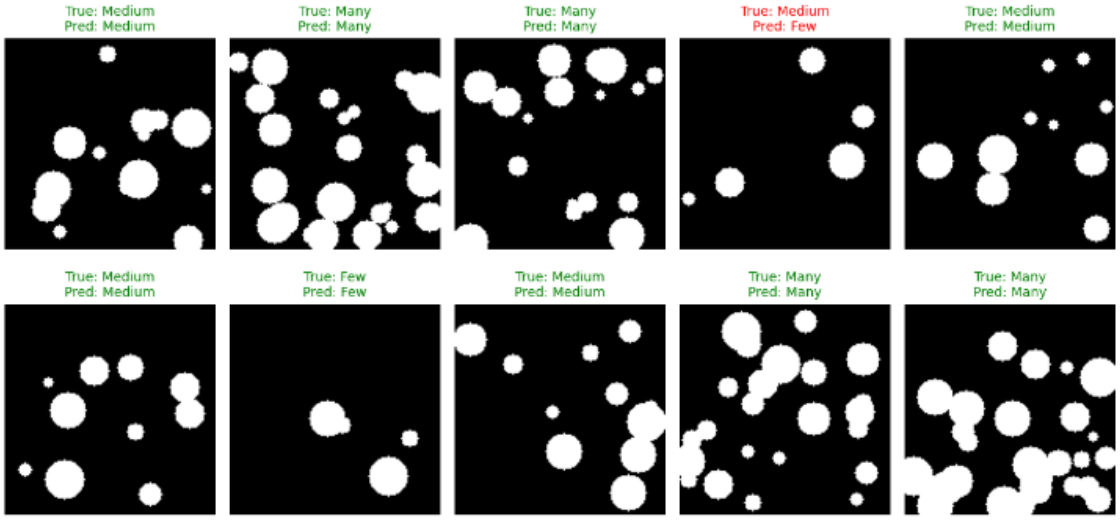

Figure 5.3: Sample predictions from the test set in Experiment 1. Correct predictions are shown in green, while misclassifications are highlighted in red, mostly involving Medium class confusion.

5.0.2 Experiment 2: Improved CNN on Dot Pattern Dataset (AdamW + Dropout)

Experiment 2 investigates whether introducing the AdamW optimizer and dropout regularization improves model generalization compared to the baseline CNN. These adjustments are expected to stabilize training and enhance robustness against overfitting.

Table 5.3 reports the final test set precision, recall, F1-score and support for each numerosity class. Compared to Experiment 1, all metrics show noticeable improvements, particularly for the "Medium" class, which was previously the hardest to classify.

Table 5.3: Final Test Set Accuracy and Per-Class Metrics for Experiment 2 (CNN with AdamW and Dropout). CNN achieves better generalization with AdamW optimizer and dropout regularization.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Few	0.92	0.91	0.91	122
Medium	0.85	**0.90**	0.88	277
Many	**0.95**	0.91	**0.93**	351
Average	**0.91**	**0.91**	**0.91**	**750**

Table 5.4 summarizes the overall test set accuracy, reaching **90.53%**, which represents an improvement over Experiment 1.

Table 5.4: Overall Test Accuracy for Experiment 2 (CNN with AdamW and Dropout).

Metric	**Value**
Final Test Accuracy	**90.53%**

Figure 5.4 presents the training curves. The left plot shows the training and validation loss decreasing over epochs, while the right plot displays validation accuracy trends. Compared to Experiment 1, the validation loss is lower and validation accuracy is notably more stable across epochs.

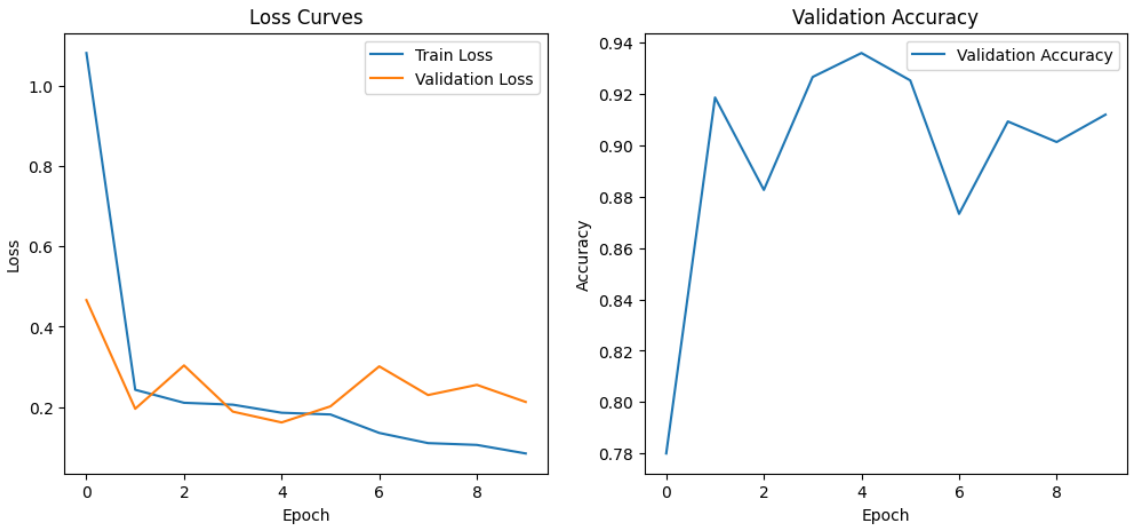

Figure 5.4: Training curves for Experiment 2 (CNN with AdamW and Dropout). Left: Training vs Validation Loss. Right: Validation Accuracy. Compared to Experiment 1, validation loss is lower and validation accuracy is more stable.

Figure 5.5 shows the confusion matrix for Experiment 2. The matrix values are percentages. Slight confusion between "Medium" and its neighboring classes persists, but diagonal dominance is strong, indicating improved class discrimination.

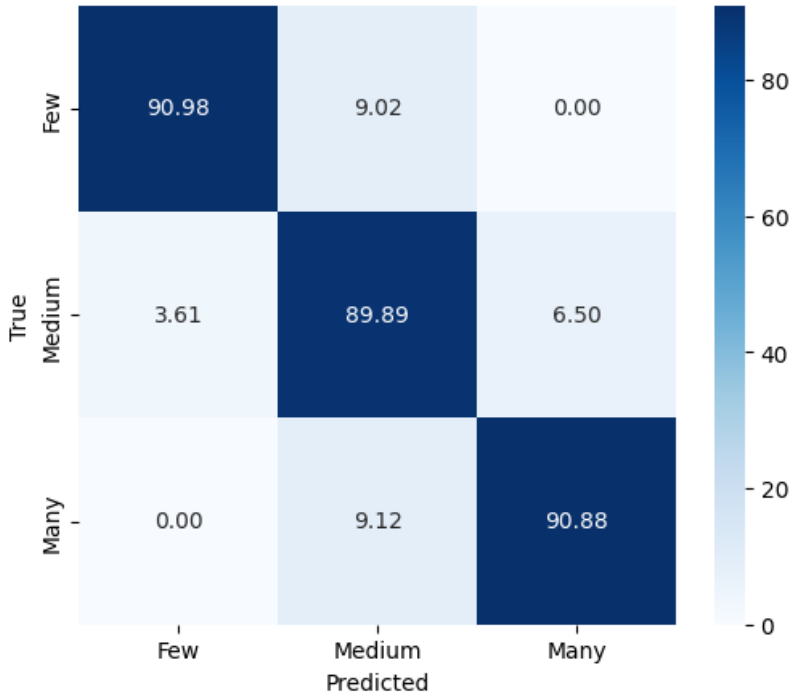

Figure 5.5: Confusion Matrix for Experiment 2. Slight confusion between "Medium" and adjacent classes persists, but diagonal dominance is strong, reflecting improved classification. Values are percentages.

Finally, Figure 5.6 presents sample predictions from the test set. Correct pre-

dictions are marked in green, while errors are highlighted in red. Visual inspection confirms the improved robustness and accuracy achieved through AdamW optimization and dropout regularization.

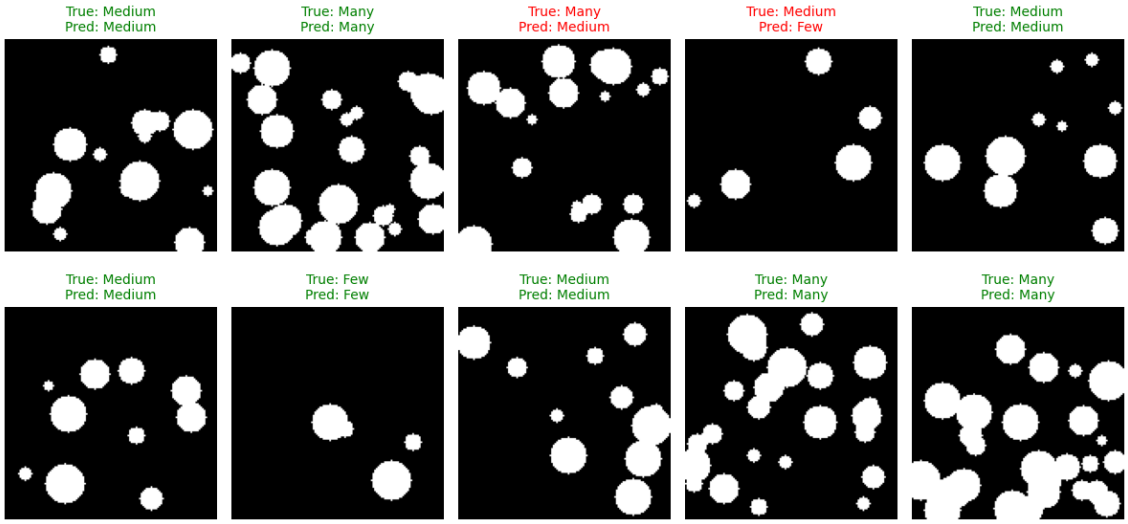

Figure 5.6: Sample predictions from the Experiment 2 test set. Correct predictions are shown in green text, misclassifications in red text. Visual inspection confirms the improved model robustness.

5.0.3 Experiment 3: CNN with Larger Batch Size, Dropout and Weight Decay

This experiment investigates the effect of using a larger batch size (128), stronger dropout (0.4) and weight decay (5×10^{-4}) to improve generalization on the dot dataset. The objective is to observe whether such regularization can help the model avoid overfitting and perform better across classes.

Table 5.5 presents the final test set precision, recall and F1-scores for each class. The "Few" class achieves near-perfect recall, while "Medium" remains more challenging, with a lower precision and F1-score. The average scores across all classes reflect solid overall performance.

Table 5.5: Final Test Set Accuracy and Per-Class Metrics for Experiment 3 (CNN with larger batch size, dropout and weight decay). The "Few" class achieves near-perfect recall, while "Medium" remains more challenging.

Class	Precision	Recall	F1-Score	Support
Few	0.85	**0.98**	0.91	122
Medium	0.84	0.89	0.87	277
Many	**0.97**	0.87	**0.92**	351
Average	**0.89**	**0.91**	**0.90**	**750**

5.0.4 Experiment 4: CNN with Tuned Regularization and Learning Rate

In this experiment, we evaluate the effect of fine-tuned hyperparameters, including a learning rate of 0.0002, dropout of 0.3 and a larger batch size of 256. This setup aims to improve generalization by encouraging regularization and smoother gradient updates. Table 5.6 presents the per-class precision, recall and F1-score, while Table 5.7 shows the overall test set accuracy.

Table 5.6: Final Test Set Accuracy and Per-Class Metrics for Experiment 4 (CNN with tuned learning rate, dropout and batch size). Experiment 4 shows stronger balancing across classes, with "Few" maintaining near-perfect recall.

Class	Precision	Recall	F1-Score	Support
Few	0.85	**0.98**	0.91	122
Medium	0.86	0.86	0.86	277
Many	**0.95**	0.90	**0.92**	351
Average	**0.89**	**0.91**	**0.90**	**750**

Table 5.7: Overall Test Accuracy for Experiment 4 (CNN: LR = 0.0002, Dropout = 0.3, Batch Size = 256).

Metric	Value
Final Test Accuracy	**89.87%**

Figure 5.7 shows the loss and validation accuracy curves. Compared to earlier configurations, the training and validation losses converge more tightly and validation accuracy reaches 93.07% around epoch 19, indicating stronger optimization stability.

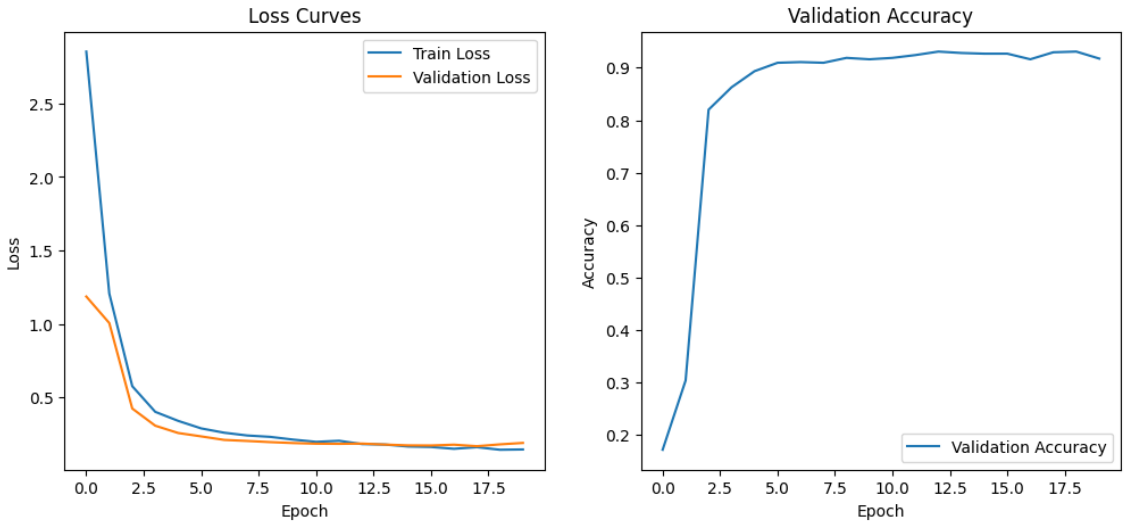

Figure 5.7: Training curves for Experiment 4 (CNN with tuned hyperparameters). Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. Validation accuracy reached 93.07% around epoch 19, indicating stronger convergence compared to earlier experiments.

Figure 5.8 shows the confusion matrix. The "Few" class maintains near-perfect recall. The "Medium" and "Many" classes are more clearly separated than in previous experiments, with stronger diagonal dominance.

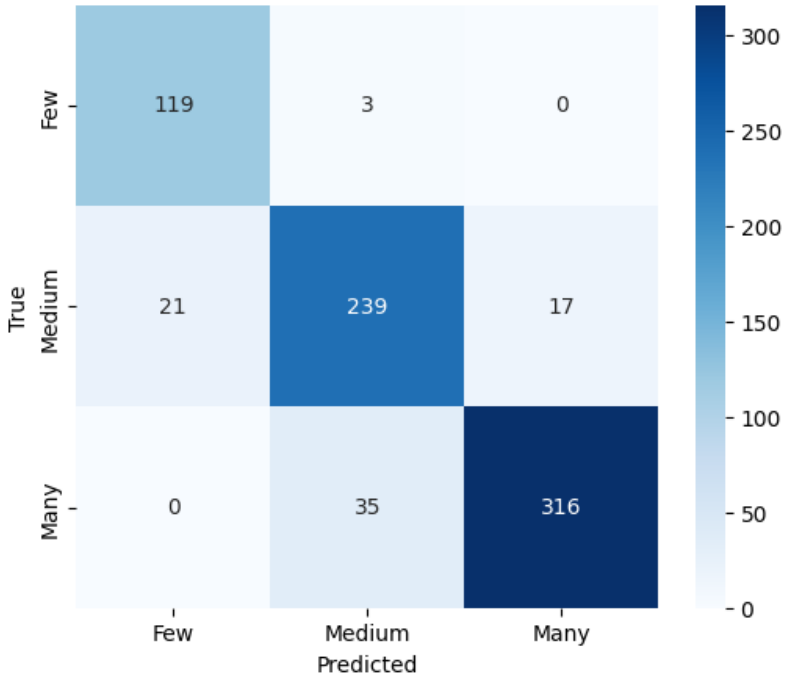

Figure 5.8: Confusion Matrix for Experiment 4 (CNN on Dot Dataset). Class "Few" maintains near-perfect recall, with improved precision and recall balance in "Medium" and "Many" categories. Values are shown as percentages.

Figure 5.9 presents qualitative results using sample predictions. Green labels

mark correct predictions, while red ones indicate errors. Compared to prior results, fewer red-labeled examples are observed in "Medium" versus "Many," confirming better inter-class separation.

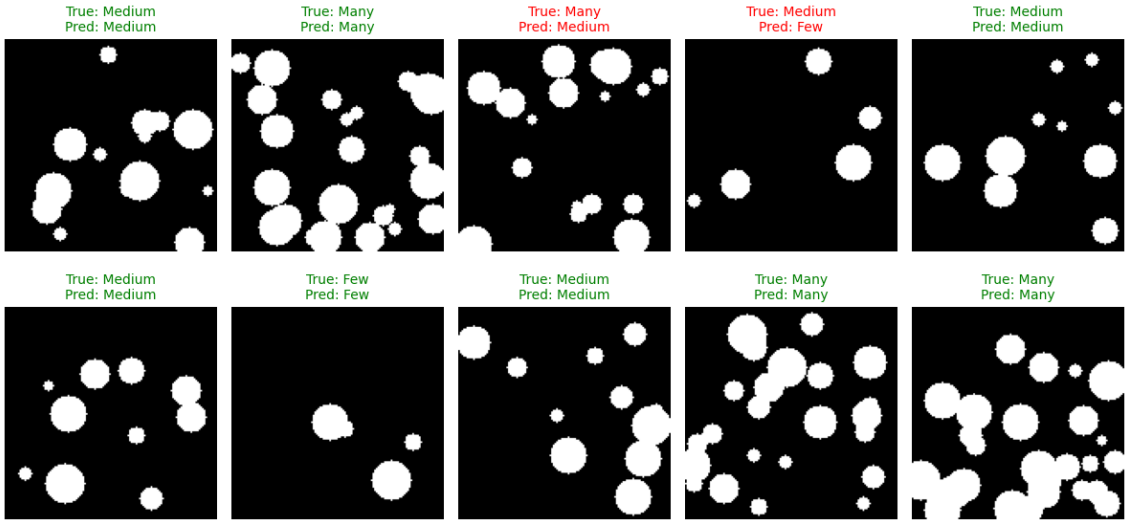

Figure 5.9: Sample predictions for Experiment 4. Correctly classified samples are indicated in green, while misclassifications are shown in red. Visualizations reveal further improvement in separating "Medium" from "Many".

5.0.5 Experiment 5: CNN + Transformer Hybrid Architecture on Dot Dataset

This experiment evaluates a hybrid CNN + Transformer architecture on the dot pattern dataset. As illustrated in Table 5.8, the model achieved consistently high scores across all classes, with a final test accuracy of 93.33% (Table 5.9). This is the highest among all previous experiments.

Table 5.8: Final Test Set Accuracy and Per-Class Metrics for Experiment 5 (CNN + Transformer on Dot Dataset). The hybrid model achieves consistently high precision, recall and F1-scores across all classes, demonstrating strong generalization.

Class	Precision	Recall	F1-Score	Support
Few	0.94	**0.97**	0.95	122
Medium	0.93	0.89	0.91	277
Many	**0.94**	0.95	**0.95**	351
Average	**0.94**	**0.94**	**0.94**	**750**

Table 5.9: Overall Test Set Accuracy for Experiment 5 (CNN + Transformer on Dot Dataset).

Metric	Value
Final Test Accuracy	**93.33%**

Figure 5.10 presents the training and validation metrics. The loss curves show quick convergence with minimal overfitting. Validation accuracy remains consistently high throughout training.

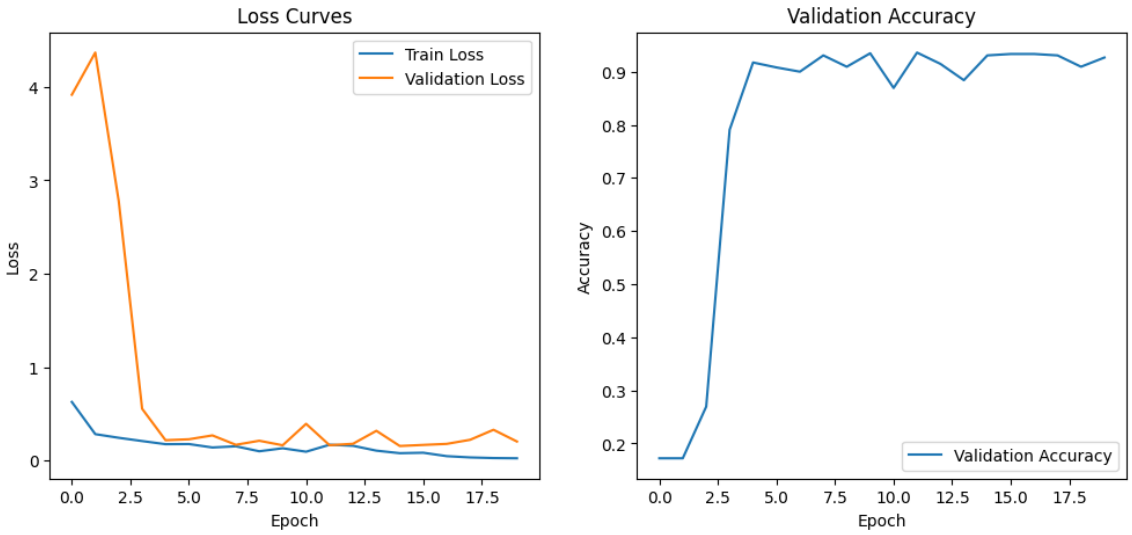

Figure 5.10: Training metrics for Experiment 5 (CNN + Transformer on Dot Dataset). Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. The model shows excellent convergence with high final validation accuracy and minimal overfitting.

Figure 5.11 shows the confusion matrix. The model maintains over 87% correct classification across all classes, with only minor confusion between "Medium" and "Many" likely due to their visual similarity.

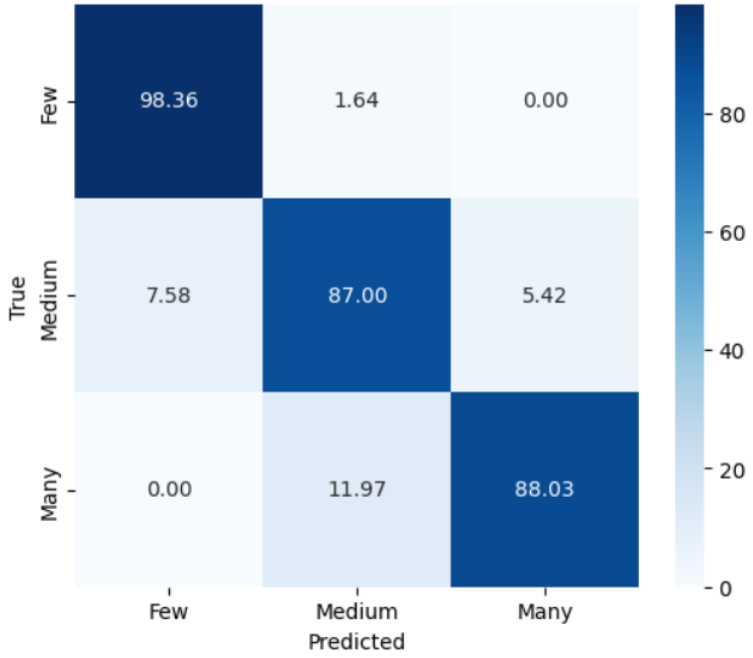

Figure 5.11: Confusion Matrix for Experiment 5 (CNN + Transformer). Excellent recall and precision observed across all numerosity classes, with only minor confusion between "Medium" and "Many". Values normalized to percentage scale.

Finally, Figure 5.12 showcases sample predictions. Correct predictions are labeled in green, while red text indicates misclassified samples. The model demonstrates strong alignment with ground truth, particularly in distinguishing "Few" and "Many."

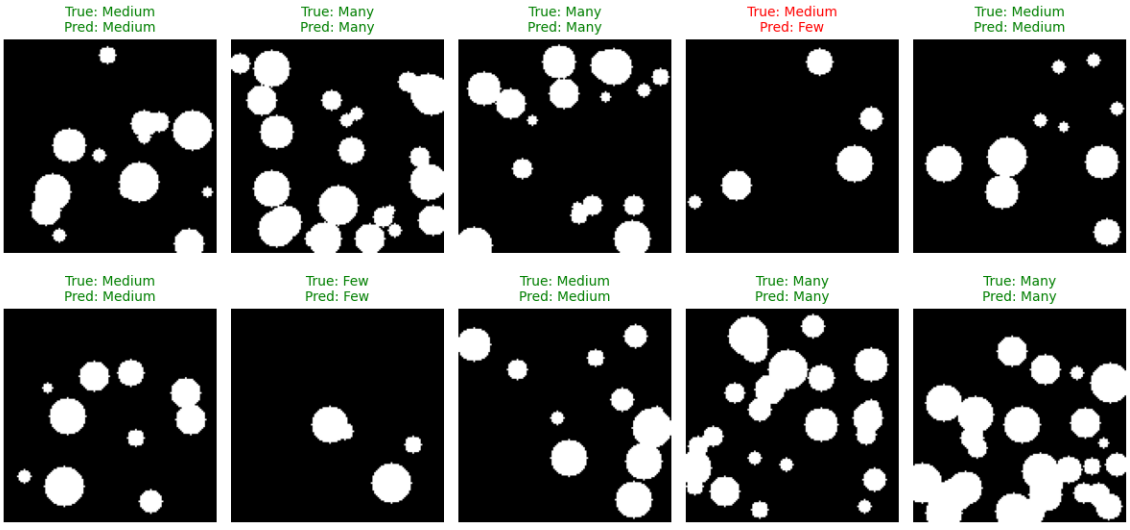

Figure 5.12: Sample predictions from Experiment 5 test set. Correct predictions are labeled in green; misclassifications are labeled in red. Strong agreement between ground truth and predictions.

5.0.6 Comparison of Dot Dataset Experiments (Experiments 1–5)

Table 5.10 presents a consolidated comparison of Experiments 1 through 5 conducted on the Dot Dataset. Each row corresponds to a different model configuration, showing the key architectural or training changes and the resulting test accuracy.

The baseline CNN (Experiment 1) already achieved competitive performance with 89.73% accuracy. Subsequent improvements such as dropout regularization (Experiment 2) and optimized learning rates and batch sizes (Experiments 3–4) provided marginal boosts. However, the most significant gain came from the CNN + Transformer hybrid architecture (Experiment 5), which achieved the highest accuracy of **93.33%**, clearly outperforming the CNN-only configurations.

Table 5.10: Comparison of Experiments 1–5 on the Dot Dataset. Transformer-based model (Experiment 5) achieved the highest test accuracy.

Exp.	**Model**	**Key Hyperparameters**	**Accuracy (%)**
1	CNN (Baseline)	Adam optimizer, Batch Size 64, No Dropout	89.73
2	CNN (Improved)	AdamW optimizer, Batch Size 64, Dropout 0.3	90.53
3	CNN (Larger Batch)	AdamW optimizer, Batch Size 128, Dropout 0.4, Weight Decay 5×10^{-4}	89.73
4	CNN (Tuned)	AdamW optimizer, Batch Size 256, Dropout 0.3, Learning Rate 0.0002	89.87
5	CNN + Transformer	AdamW optimizer, Batch Size 256, Dropout 0.4, Learning Rate 0.0001	**93.33**

5.0.7 Experiment 6: CNN on Silhouette Dataset

Experiment 6 evaluates whether a CNN trained from scratch on the silhouette dataset can effectively learn abstract numerosity representations despite the added visual complexity introduced by semantically meaningful shapes. Unlike dot patterns, these inputs feature diverse object silhouettes (e.g., tools, animals) with variations in contour and internal complexity.

Table 5.11 summarizes the classification performance for each numerosity class. The CNN achieved high recall for the "Many" category (96%), indicating it successfully captured larger object groupings. However, it struggled with the "Medium"

class, yielding the lowest precision and F1-score, likely due to higher intra-class variability in this mid-range category.

The final test accuracy was 79.33% (Table 5.12), showing a moderate drop from dot-based performance, consistent with the added abstraction challenge.

Figure 5.13 shows the training and validation curves. Both training and validation loss decreased steadily and validation accuracy plateaued above 75%, suggesting reasonable convergence with minimal overfitting.

The confusion matrix in Figure 5.14 shows the dominant confusion centered around the "Medium" class, which overlaps with both "Few" and "Many" in pixel density and shape diversity. Diagonal dominance, however, confirms the model's ability to distinguish "Few" and "Many" cases effectively.

Visual examples in Figure 5.15 further illustrate the model's predictions. Misclassifications, highlighted in red, mostly involve borderline cases in the "Medium" category, while predictions for extreme classes appear highly reliable.

This experiment demonstrates that a CNN can generalize numerosity categorization to perceptually diverse object silhouettes, although performance degrades relative to the simpler dot domain.

Table 5.11: Final Test Set Accuracy and Per-Class Metrics for Experiment 6 (CNN on Silhouette Dataset). The CNN shows strong performance on the "Many" class, but struggles moderately on the "Medium" class.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Few	0.90	0.73	0.81	71
Medium	0.76	0.59	0.66	156
Many	**0.79**	**0.96**	**0.86**	223
Average	**0.82**	**0.76**	**0.78**	**450**

Table 5.12: Overall Test Set Accuracy for Experiment 6 (CNN on Silhouette Dataset).

Metric	**Value**
Final Test Accuracy	**79.33%**

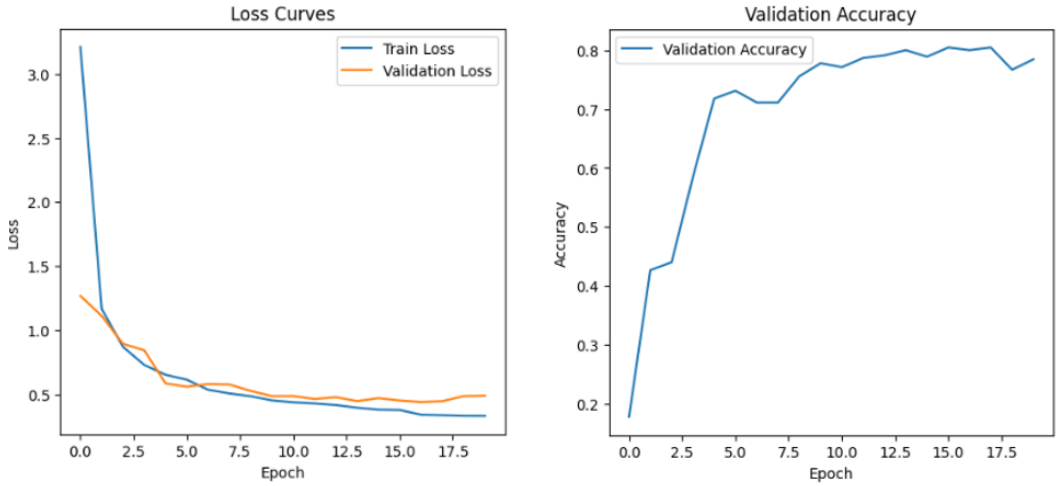

Figure 5.13: Training and validation metrics for Experiment 6. Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. Model converged moderately well on the more complex silhouette-based data.

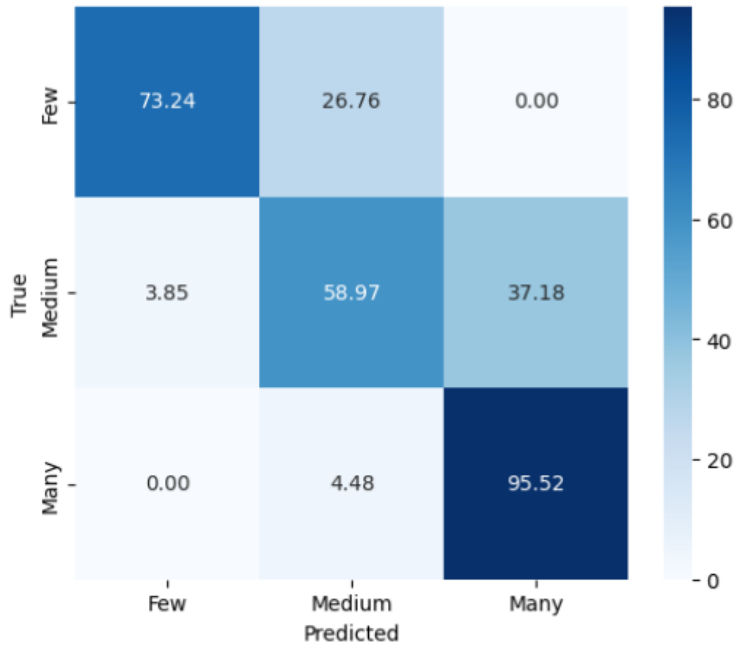

Figure 5.14: Confusion Matrix for Experiment 6. The "Many" class achieved excellent recall (96%), while the "Medium" category saw more confusion, indicating greater difficulty in distinguishing intermediate numerosity.

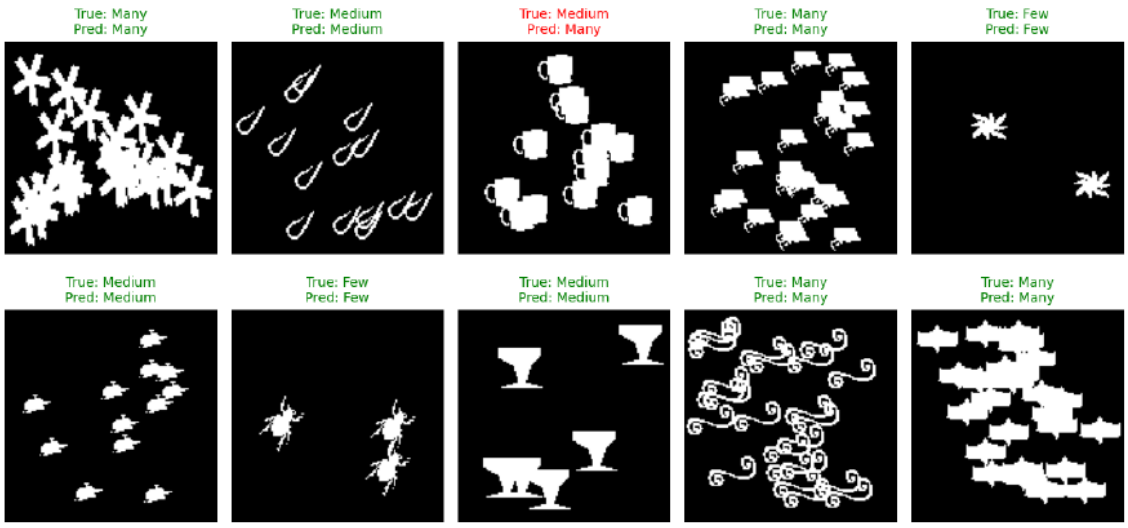

Figure 5.15: Sample predictions from the test set in Experiment 6. Correct predictions are shown in green, while misclassifications are highlighted in red.

5.0.8 Experiment 7: CNN + Transformer on Silhouette Dataset

Experiment 7 investigates how well the hybrid CNN + Transformer model performs on the more visually complex silhouette dataset. This dataset introduces intra-class variability in shape, semantics and positioning, making it a more realistic benchmark for generalization. The model configuration uses a tuned learning rate of 0.0001, a batch size of 256 and a dropout rate of 0.4.

Table 5.13 presents per-class precision, recall and F1-score metrics. The CNN + Transformer model performs best on the "Few" and "Many" categories, achieving F1-scores above 0.85. However, the "Medium" class remains the most challenging, with a balanced but lower performance.

Table 5.13: Final Test Set Accuracy and Per-Class Metrics for Experiment 7 (CNN + Transformer on Silhouettes). The model achieved strong performance, particularly on "Few" and "Many" classes, while "Medium" remained the most challenging.

Class	Precision	Recall	F1-Score	Support
Few	0.83	**0.89**	0.86	71
Medium	0.74	0.75	0.75	156
Many	**0.88**	0.85	**0.87**	223
Average	**0.82**	**0.83**	**0.83**	**450**

Table 5.14 shows the overall test accuracy of 82.22%, the highest observed on the silhouette dataset. This reinforces the Transformer's ability to abstract over semantic variability in shapes.

Table 5.14: Overall Test Set Accuracy for Experiment 7 (CNN + Transformer on Silhouette Dataset).

Metric	Value
Final Test Accuracy	**82.22%**

Figure 5.16 shows training and validation metrics. The training loss decreases consistently and the validation accuracy reaches above 80% by epoch 10, showing good generalization with minimal overfitting.

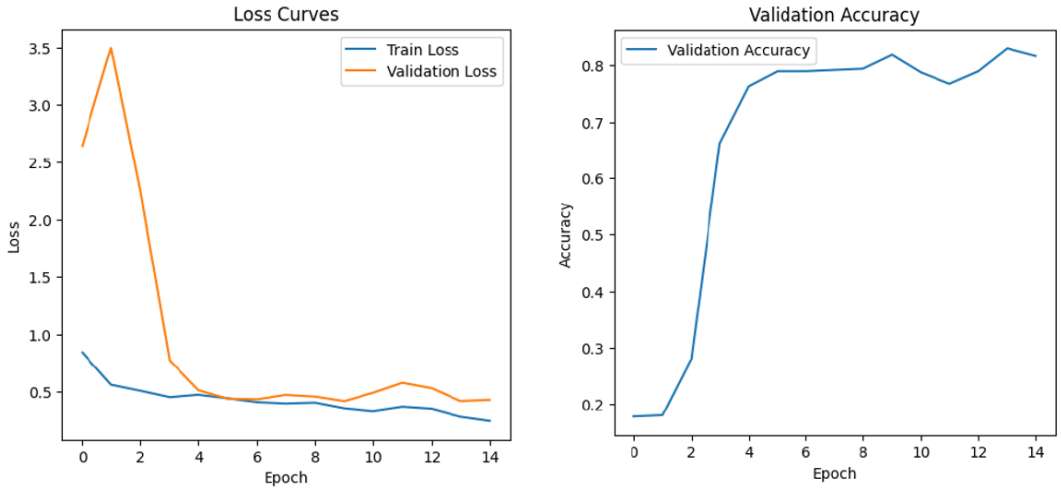

Figure 5.16: Training metrics for Experiment 7. Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. The Transformer model achieved smooth convergence and good validation performance without major overfitting.

Figure 5.17 displays the confusion matrix. Most "Few" and "Many" examples are correctly classified, while the "Medium" class is often confused with its neighbors, a trend consistent with earlier experiments.

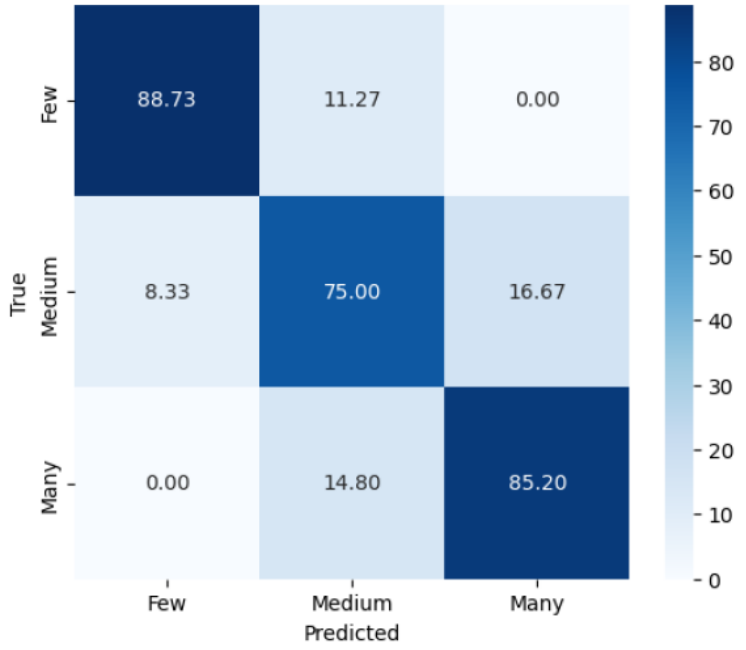

Figure 5.17: Confusion Matrix for Experiment 7. Excellent precision and recall for "Few" and "Many" classes. Misclassifications mostly involved the "Medium" class.

Figure 5.18 shows sample predictions. Green-labeled images denote correct classifications, while red indicates misclassifications. Visual review confirms strong qualitative alignment with numerical performance.

Figure 5.18: Sample predictions from the test set in Experiment 7. Correct predictions are shown in green; incorrect predictions are shown in red.

5.0.9 Comparison of Silhouette Dataset Experiments (Experiments 6–7)

This section compares the performance of the CNN-only model (Experiment 6) and the CNN + Transformer hybrid model (Experiment 7) on the silhouette dataset, which consists of complex and semantically diverse object shapes.

Table 5.15 summarizes the key hyperparameters and final test accuracy for both models. The CNN + Transformer model (Experiment 7) achieved an accuracy of 82.22%, slightly higher than the CNN model's 79.33% (Experiment 6). This suggests that the Transformer's global attention mechanism contributed to improved generalization on complex visual input.

Both experiments used the same batch size and AdamW optimizer, but differed in learning rate and dropout configuration. These results demonstrate the benefits of using hybrid architectures when handling semantically richer inputs, such as silhouettes from unrelated object categories.

Table 5.15: Comparison of Experiments 6–7 on the Silhouette Dataset. CNN + Transformer model (Experiment 7) slightly outperformed the pure CNN model.

Exp.	**Model**	**Key Hyperparameters**	**Accuracy (%)**
6	CNN	AdamW optimizer, Batch Size 256, Dropout 0.3, Learning Rate 0.0002	79.33
7	CNN + Transformer	AdamW optimizer, Batch Size 256, Dropout 0.4, Learning Rate 0.0001	**82.22**

5.0.10 Generalization to Dot Dataset Variants

To assess whether models rely on abstract numerosity or low-level visual features, the dot-trained CNN and CNN + Transformer models were evaluated on three controlled variants: **Shape**, **Occlusion** and **Clustered**. These variants distort object identity, spatial arrangement and visibility respectively, without changing the underlying numerosity categories.

Clustered Variant

This experiment evaluates how well models trained on standard dot patterns can generalize to images where dots are tightly clustered. Although the number of items remains constant within each class, their dense spatial arrangement challenges the

model's ability to distinguish individual elements, a well-documented limitation in both human and artificial perception.

CNN + Transformer Model

The CNN + Transformer model exhibits partial robustness to spatial compression. As shown in Figure 5.19, while some "Medium" samples are correctly classified, many "Many" instances are misidentified as "Medium" or "Few." This reflects a breakdown in generalization, despite the model's global attention mechanism.

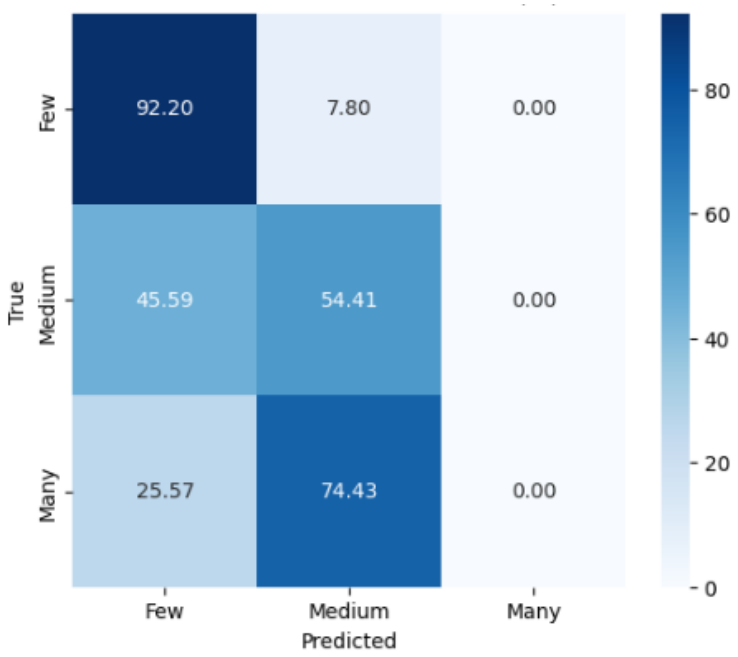

Figure 5.19: Confusion Matrix for Clustered Variant (CNN + Transformer). Performance breaks down under clustered arrangements, with many samples misclassified as "Few".

Sample predictions in Figure 5.20 show the same underestimation trend. High-density arrangements appear to confuse the model, which often interprets them as lower-count classes. These findings suggest that the attention-based architecture still struggles to abstract numerosity under extreme clustering.

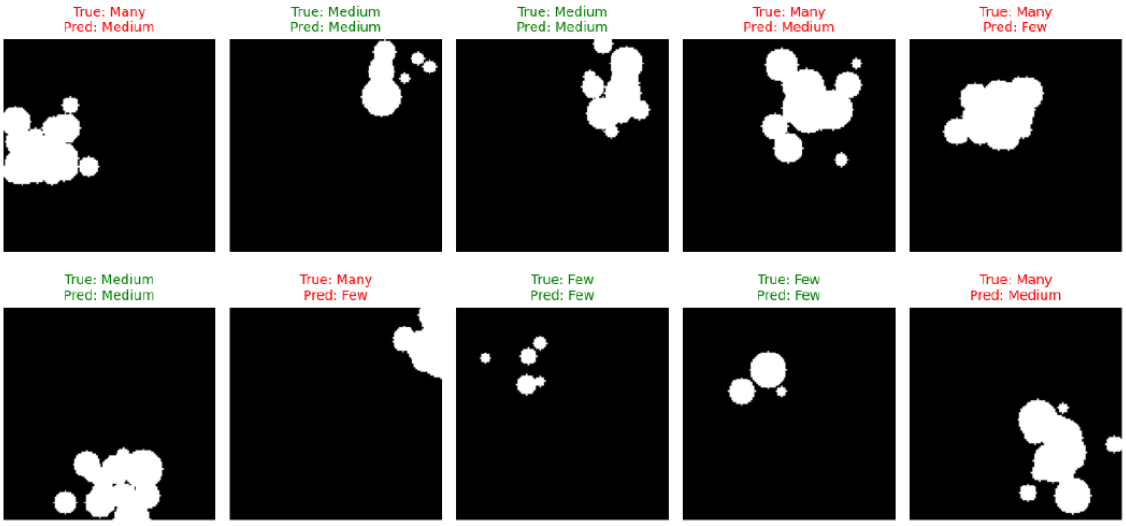

Figure 5.20: Sample Predictions for Clustered Variant (CNN + Transformer). High visual density leads to frequent misclassifications across classes.

CNN Only Model

The CNN-only model performs notably worse. Figure 5.21 shows that nearly all samples, even those truly belonging to the "Many" category, are predicted as "Few." This suggests the model is relying heavily on overall white pixel density rather than discrete object individuation.

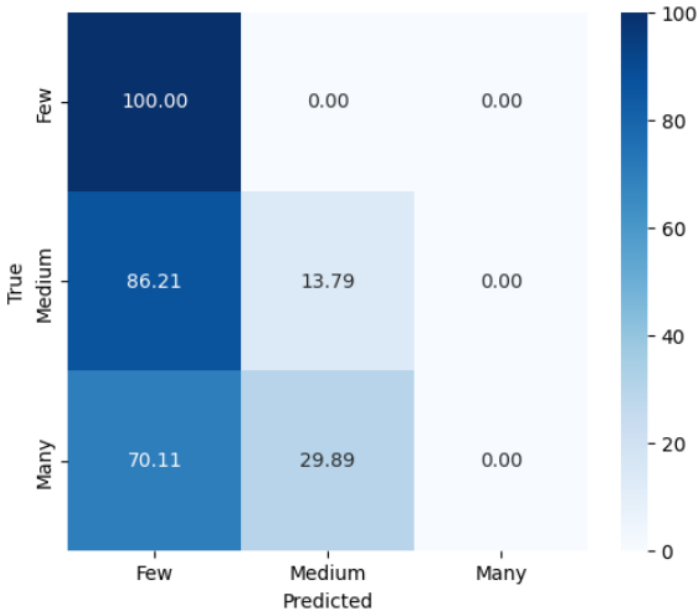

Figure 5.21: Confusion Matrix for Clustered Variant (CNN Only). Most samples were predicted as "Few", regardless of actual count, indicating lack of abstraction.

The prediction examples in Figure 5.22 reinforce this observation. Most samples are uniformly misclassified as "Few," regardless of actual object count. This behavior

highlights the CNN model's inability to handle tightly clustered objects, revealing its vulnerability to spatial bias and lack of abstraction.

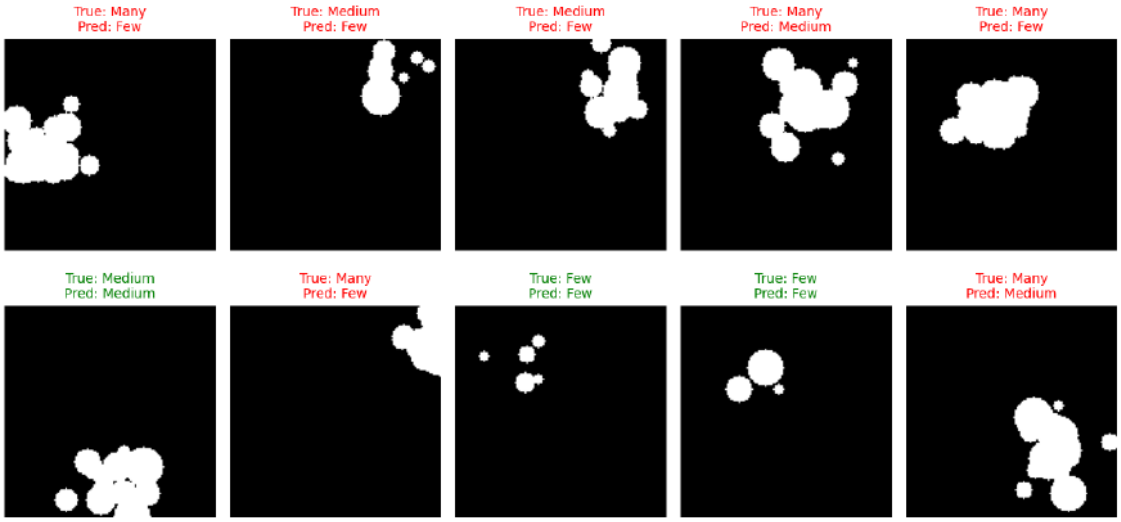

Figure 5.22: Sample Predictions for Clustered Variant (CNN Only). CNN model struggled to generalize numerosity under clustered spatial structure.

Occlusion Variant

To evaluate the robustness of numerosity abstraction under partial visibility, we tested the dot-trained CNN and CNN + Transformer models on occluded dot images. These images contained randomly placed black masks occluding parts of the white dots, while maintaining the same object count per class.

CNN + Transformer Model

Figure 5.23 shows the confusion matrix for the CNN + Transformer model on the occlusion variant. The model maintained high classification accuracy across all categories despite occlusions. Most notably, the "Many" class retained 93.85% accuracy, with minor confusion in the "Medium" category, indicating robust abstraction.

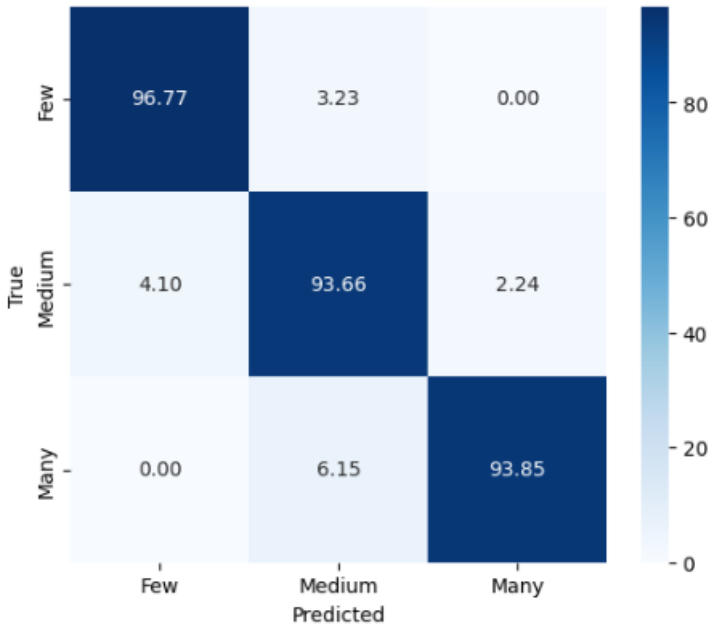

Figure 5.23: Confusion Matrix for Occlusion Variant (CNN + Transformer). Maintained strong accuracy despite partial occlusions, indicating robust numerosity abstraction.

Figure 5.24 presents sample predictions from the test set. Correct predictions (in green) dominate, with only a few misclassified samples (in red), suggesting that minor occlusions did not impact the performance of the model significantly.

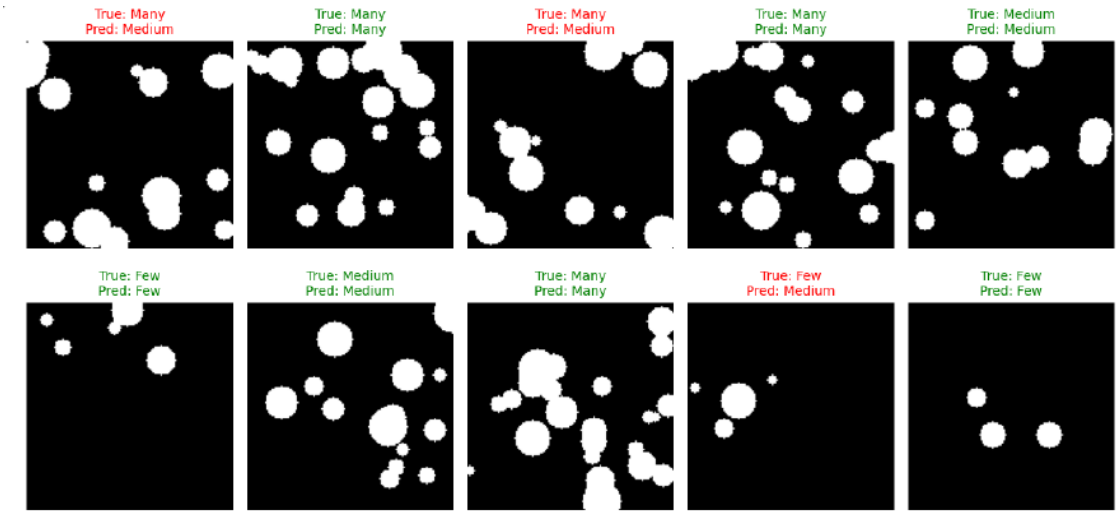

Figure 5.24: Sample Predictions for Occlusion Variant (CNN + Transformer). Minor occlusions did not significantly degrade classification performance.

CNN Only Model

For comparison, the CNN model also achieved strong performance under occlusion (Figure 5.25). While slightly lower than the transformer-enhanced model, the

CNN still classified most samples correctly, with some confusion between “Medium” and “Many”.

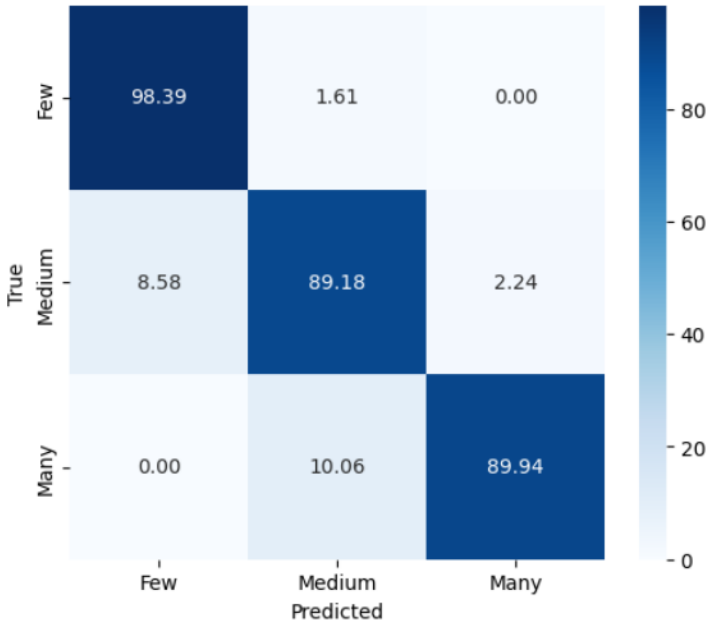

Figure 5.25: Confusion Matrix for Occlusion Variant (CNN Only). The CNN model also handled occlusions well, showing strong generalization.

Figure 5.26 shows representative predictions for the CNN model. The visual results confirm that both models generalize well to partial occlusions, with the Transformer model showing marginally stronger robustness.

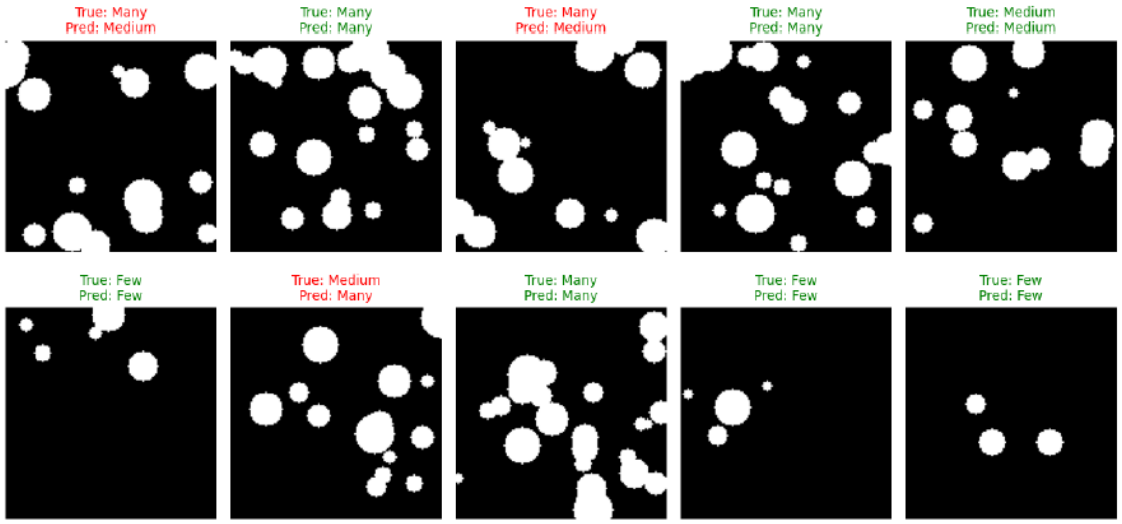

Figure 5.26: Sample Predictions for Occlusion Variant (CNN Only). The model maintained high accuracy even with partially occluded objects.

Shape Variant

To test abstraction beyond low-level object geometry, we evaluated dot-trained models on a shape-variant dataset. In this variant, dots were replaced by random geometric shapes (e.g., triangles, squares, ellipses), while maintaining the same numerosity classes. This tests whether models can generalize numerosity perception beyond specific object forms.

CNN + Transformer Model

Figure 5.27 shows the confusion matrix for the CNN + Transformer model. The model generalized well across geometric distortions, maintaining strong classification performance. The "Medium" class experienced slightly more confusion, but overall accuracy remained high, confirming that the Transformer could abstract numerosity independently of shape identity.

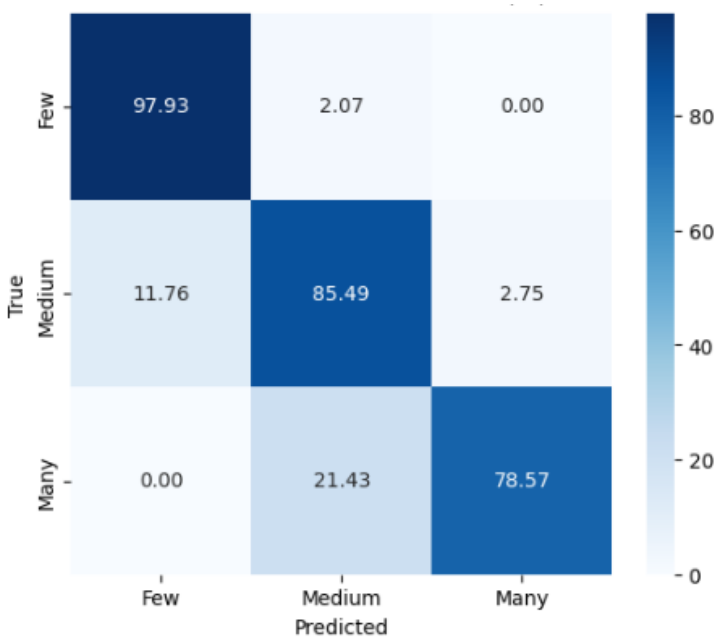

Figure 5.27: Confusion Matrix for Shape Variant (CNN + Transformer). The model generalized well across variations in object geometry, preserving numerosity perception.

Sample predictions in Figure 5.28 support this interpretation. Most predictions align well with ground truth labels, suggesting resilience to changes in shape. Errors are minimal and mostly occur in "Medium" class boundaries.

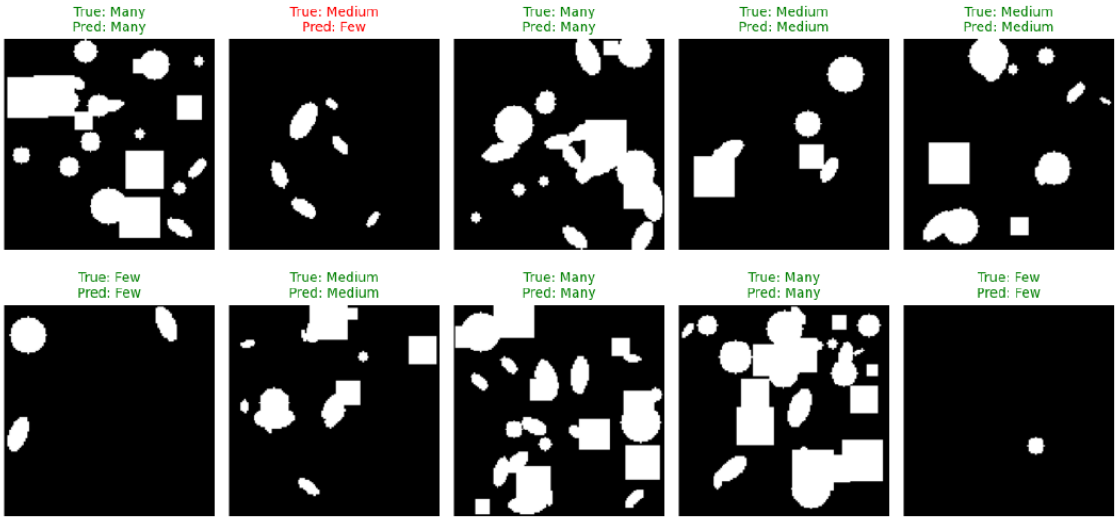

Figure 5.28: Sample Predictions for Shape Variant (CNN + Transformer). Numerosity abstraction appeared resilient to changes in object shape.

CNN Only Model

Figure 5.29 presents the CNN-only model's confusion matrix. Like the hybrid model, the CNN demonstrates strong generalization. Despite the altered object shapes, precision and recall remain high across all classes, especially "Few" and "Medium."

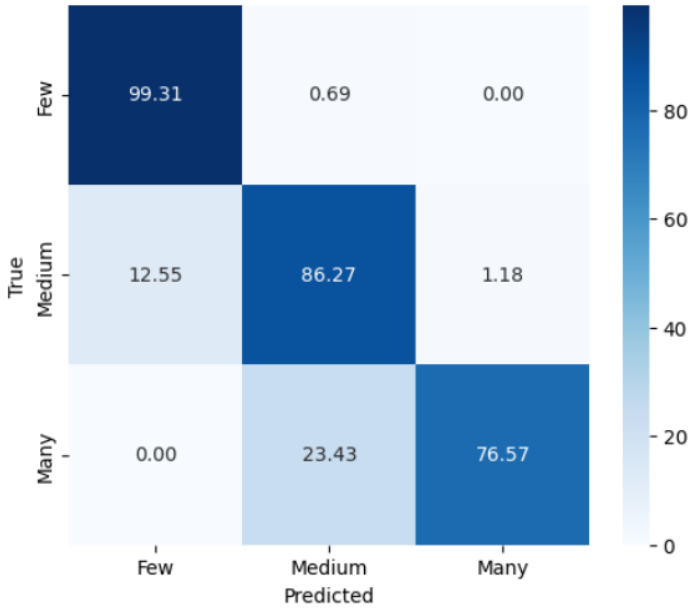

Figure 5.29: Confusion Matrix for Shape Variant (CNN Only). The CNN model maintained stable performance despite geometric distortions.

Figure 5.30 displays representative test predictions. Like the Transformer model, the CNN correctly classifies most samples, effectively ignoring shape identity and

relying on spatial or quantity-related cues.

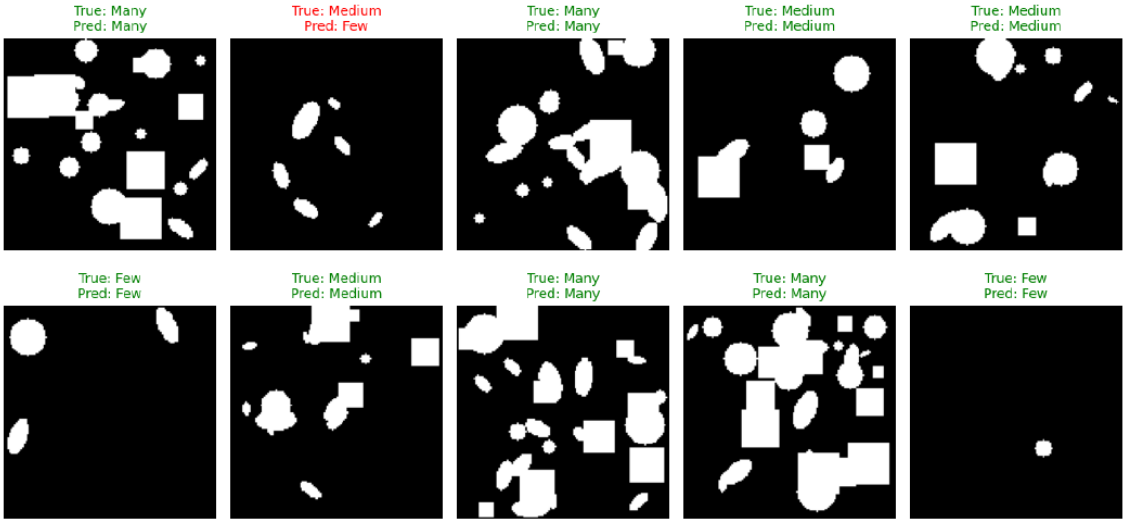

Figure 5.30: Sample Predictions for Shape Variant (CNN Only). The model effectively ignored object shape, focusing on numerosity estimation.

Summary of Variant Performance

To synthesize the findings across dot dataset variants, Table 5.16 summarizes test performance metrics for both the CNN-only and CNN + Transformer models. Metrics include test loss, accuracy, precision, recall and F1-score across Clustered, Occlusion and Shape conditions.

Table 5.16: Summary of Generalization Performance on Dot Dataset Variants. CNN + Transformer model consistently outperformed CNN-only model in Occlusion and Shape variants, but both struggled heavily under Clustered arrangements.

Dataset Variation	Model	Test Loss	Accuracy (%)	Precision	Recall	F1-Score
Clustered	CNN + Transformer	4.9197	**36.27**	**0.24**	**0.49**	**0.32**
Clustered	CNN only	5.4585	23.60	0.16	0.38	0.19
Occlusion	CNN + Transformer	0.2119	**94.27**	**0.93**	**0.95**	**0.94**
Occlusion	CNN only	0.2228	91.07	0.90	0.93	0.91
Shape	CNN + Transformer	0.6276	**84.67**	**0.85**	0.87	**0.85**
Shape	CNN only	0.3675	84.27	0.84	**0.87**	**0.85**

The table highlights a clear trend: both models performed strongly in the Occlusion and Shape variants, with the CNN + Transformer model showing a slight edge in overall accuracy and precision. This suggests that both architectures are capable of abstracting numerosity even under partial visibility and geometric distortion.

However, performance on the Clustered variant reveals a significant breakdown. The CNN-only model misclassified nearly all inputs, while the Transformer-enhanced model managed slightly better but still suffered from poor precision and over-reliance on predicting the “Few” class. These results confirm the models’ shared difficulty in handling tightly packed object layouts, where individuation becomes perceptually challenging.

In summary, while both models generalize well under mild visual variation, they remain sensitive to spatial density distortions, an area requiring architectural and dataset-level improvements in future work.

5.0.11 Cross-Modality Generalization: Dot to Silhouette

The generalization ability of models trained on dot patterns was evaluated by directly testing them on the silhouette dataset without fine-tuning. Results indicate poor transfer, with both CNN and CNN + Transformer models overwhelmingly predicting the "Many" class regardless of input.

Table 5.17: Classification Report: Dot-Trained CNN Model Tested on Silhouette Dataset. The CNN model failed to generalize and defaulted to predicting the “Many” class.

Class	Precision	Recall	F1-Score	Support
Few	0.00	0.00	0.00	71
Medium	0.00	0.00	0.00	156
Many	**0.50**	**1.00**	**0.66**	223

Table 5.18: Classification Report: Dot-Trained CNN + Transformer Model Tested on Silhouette Dataset. The hybrid model also failed to generalize, similarly defaulting to the “Many” class.

Class	Precision	Recall	F1-Score	Support
Few	0.00	0.00	0.00	71
Medium	0.00	0.00	0.00	156
Many	**0.50**	**1.00**	**0.66**	223

5.0.12 Cross-Domain Fine-Tuning Performance

Fine-Tuning Silhouette-Trained Models on Dot Data

This experiment investigates whether models trained on silhouette images can successfully generalize to dot patterns via fine-tuning. Table 5.19 summarizes the final performance after adaptation. Both the CNN and CNN + Transformer models reached over 91% test accuracy, with the hybrid model slightly outperforming the CNN across all metrics.

Table 5.19: Performance After Fine-Tuning Silhouette-Trained Models on Dot Data. The CNN + Transformer model slightly outperformed the CNN model across all metrics.

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN	91.60	0.91	0.92	0.92
CNN + Transformer	**92.53**	**0.92**	**0.93**	**0.93**

CNN Model

Figures 5.31–5.33 detail the performance of the fine-tuned CNN. Figure 5.31 shows that the model converges steadily during training, with validation accuracy exceeding 90% early and remaining stable.

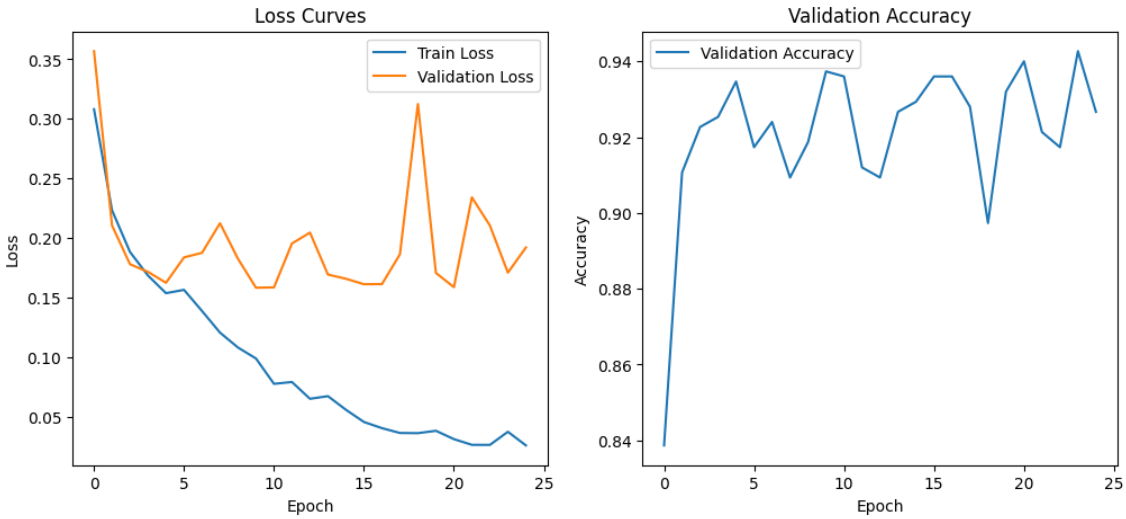

Figure 5.31: Training Metrics for Fine-Tuned CNN (Silhouette → Dot). Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. The model converged stably, achieving over 90% validation accuracy.

The confusion matrix in Figure 5.32 shows balanced performance across the three classes, with very few misclassifications and only minor confusion between

neighboring categories.

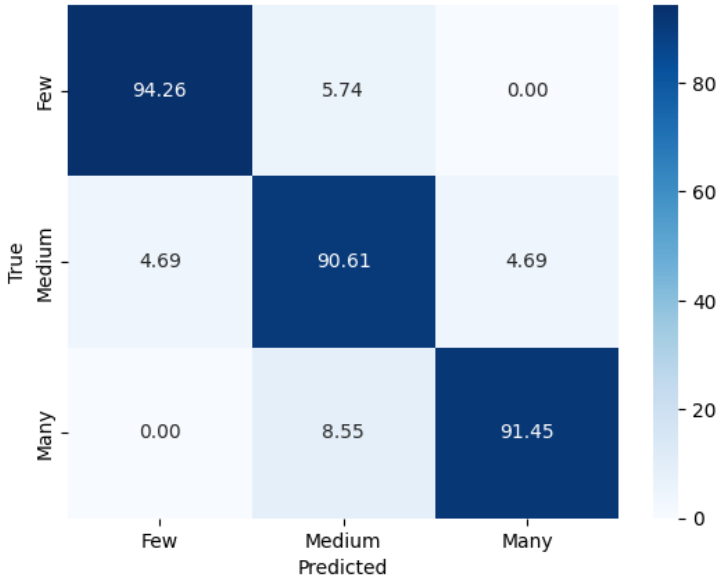

Figure 5.32: Confusion Matrix for Fine-Tuned CNN (Silhouette → Dot). High recall and precision across all classes, with only minor confusion between adjacent categories.

Sample predictions in Figure 5.33 confirm visual alignment between model output and ground truth labels. Correct predictions dominate, shown in green, with a few red labels indicating misclassification.

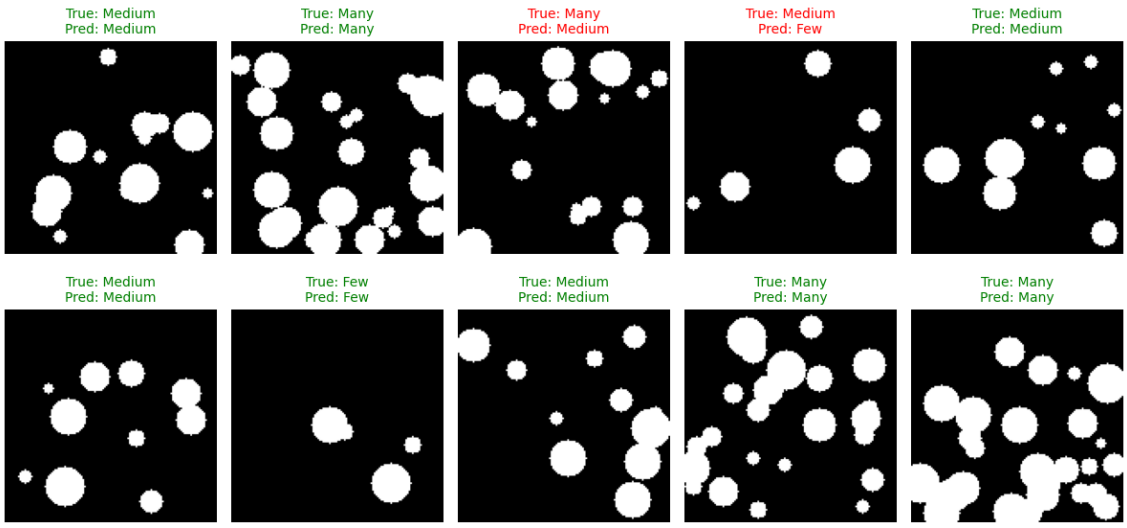

Figure 5.33: Sample Predictions for Fine-Tuned CNN (Silhouette → Dot). Correct predictions are shown in green and misclassifications are shown in red. The majority of samples were correctly classified, reflecting successful cross-domain adaptation.

CNN + Transformer Model

Figures 5.34–5.36 present the results for the CNN + Transformer model. As shown in Figure 5.34, the model exhibits smooth loss reduction and high validation accuracy throughout training.

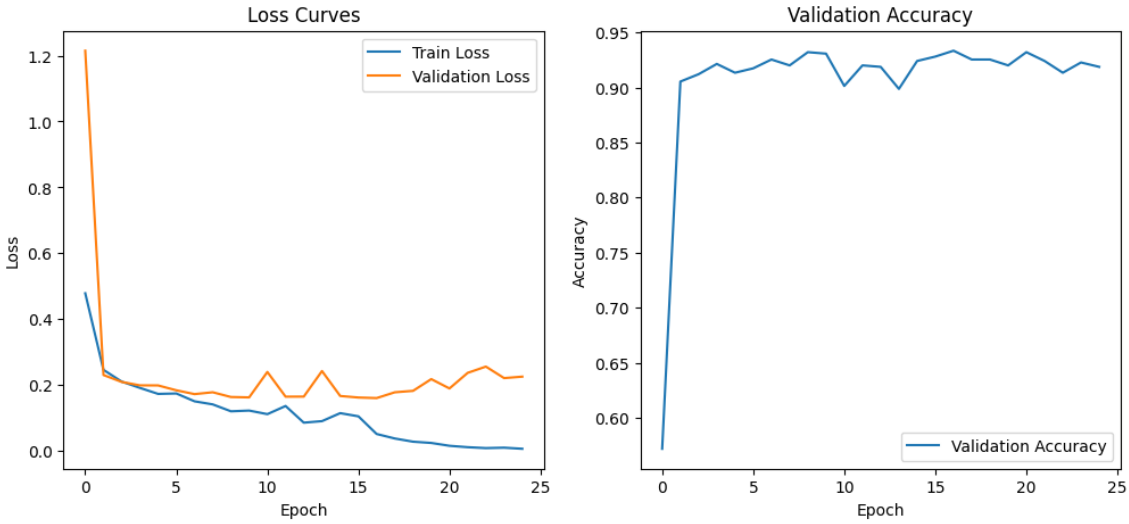

Figure 5.34: Training Metrics for Fine-Tuned CNN + Transformer (Silhouette → Dot). Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. Smooth convergence and consistently high validation accuracy were observed throughout training.

Figure 5.35 shows strong classification ability in the confusion matrix. Both "Few" and "Many" classes achieve over 95% recall, with only slight confusion in the "Medium" class.

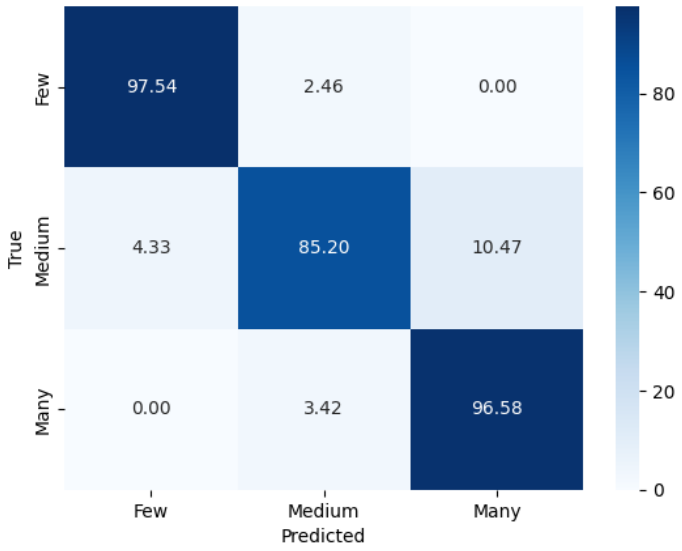

Figure 5.35: Confusion Matrix for Fine-Tuned CNN + Transformer (Silhouette → Dot). Strong classification performance across all numerosity classes, especially "Few" and "Many".

Visual samples in Figure 5.36 highlight the model's strong alignment with true labels. Errors are sparse and largely restricted to "Medium" class overlap, with excellent precision across the board.

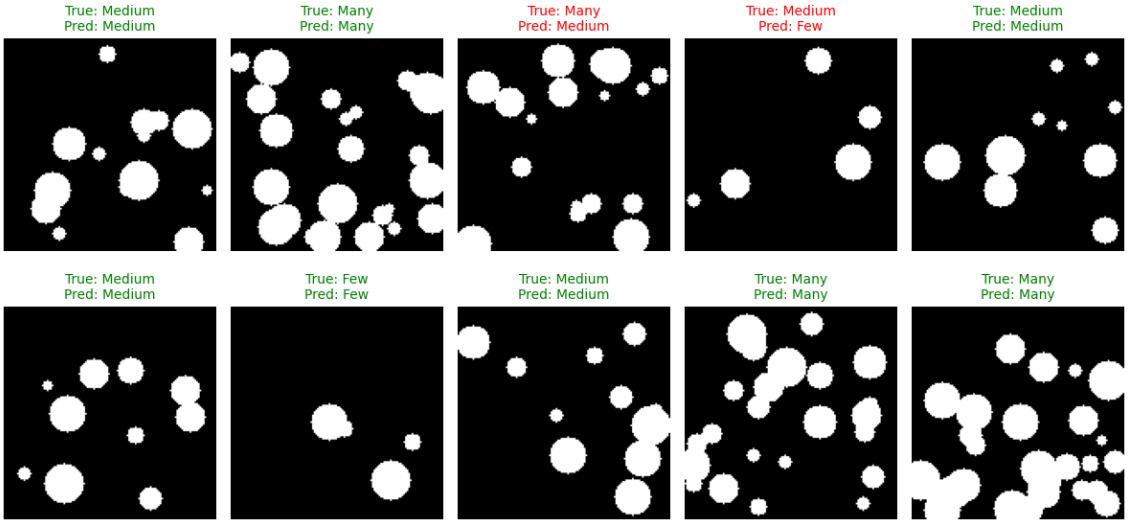

Figure 5.36: Sample Predictions from Fine-Tuned CNN + Transformer (Silhouette → Dot). High precision maintained across all categories. Correct predictions are shown in green; errors in red.

Fine-Tuning Dot-Trained Models on Silhouette Data

Fine-tuning dot-trained models on the silhouette dataset enabled evaluation of cross-domain transfer in the reverse direction. Table 5.20 summarizes the results. The CNN + Transformer model achieved slightly higher test accuracy (78.22%), while the plain CNN achieved better precision (0.78) and recall (0.80), indicating that both models adapted reasonably well, though with different tradeoffs.

Table 5.20: Performance After Fine-Tuning Dot-Trained Models on Silhouette Data. Fine-tuning significantly improved transfer ability, with CNN + Transformer achieving slightly higher overall accuracy, while CNN achieved better precision and recall.

Model	Accuracy (%)	Precision	Recall	F1-Score
CNN	76.00	**0.78**	**0.80**	0.77
CNN + Transformer	**78.22**	0.77	0.78	**0.78**

CNN Model

Training metrics for the fine-tuned CNN model are shown in Figure 5.37. The model converged smoothly with minimal overfitting, but final accuracy was lower

than that of the hybrid model. The confusion matrix in Figure 5.38 shows relatively strong performance on "Few" and "Medium", but considerable confusion between "Medium" and "Many" categories.

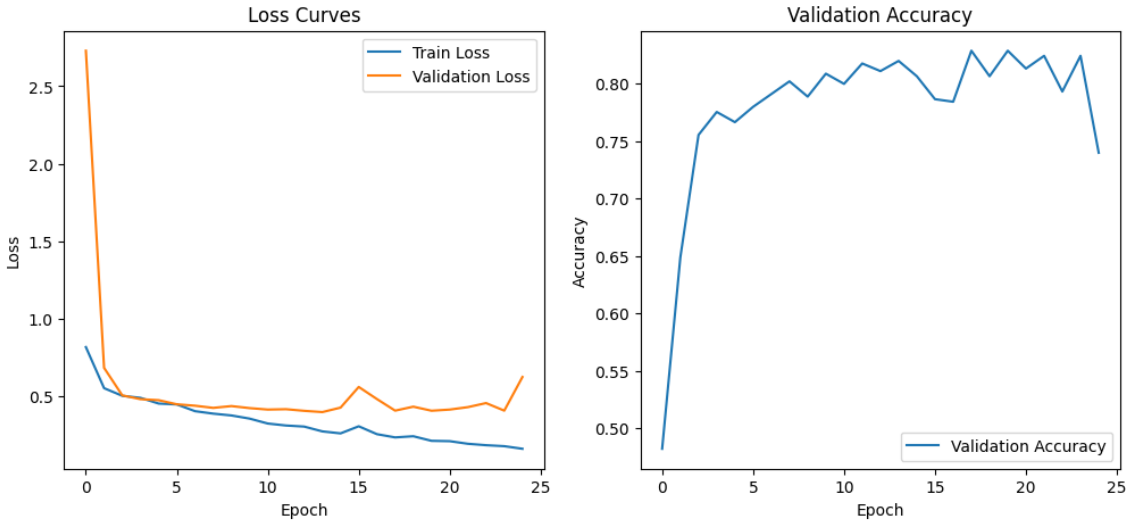

Figure 5.37: Training Metrics for Fine-Tuned CNN (Dot → Silhouette). Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. The CNN model achieved stable learning, but overall accuracy was lower than that of the hybrid model.

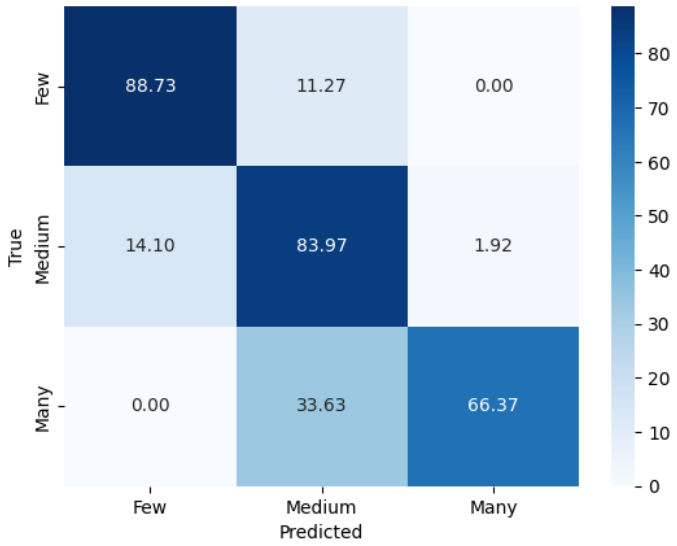

Figure 5.38: Confusion Matrix for Fine-Tuned CNN (Dot → Silhouette). While the model performed decently for "Few" and "Medium" classes, significant confusion remained between "Medium" and "Many" classes.

Qualitative predictions are shown in Figure 5.39. Most errors involved confusion between intermediate and high numerosity levels, confirming the limitations already visible in the confusion matrix.

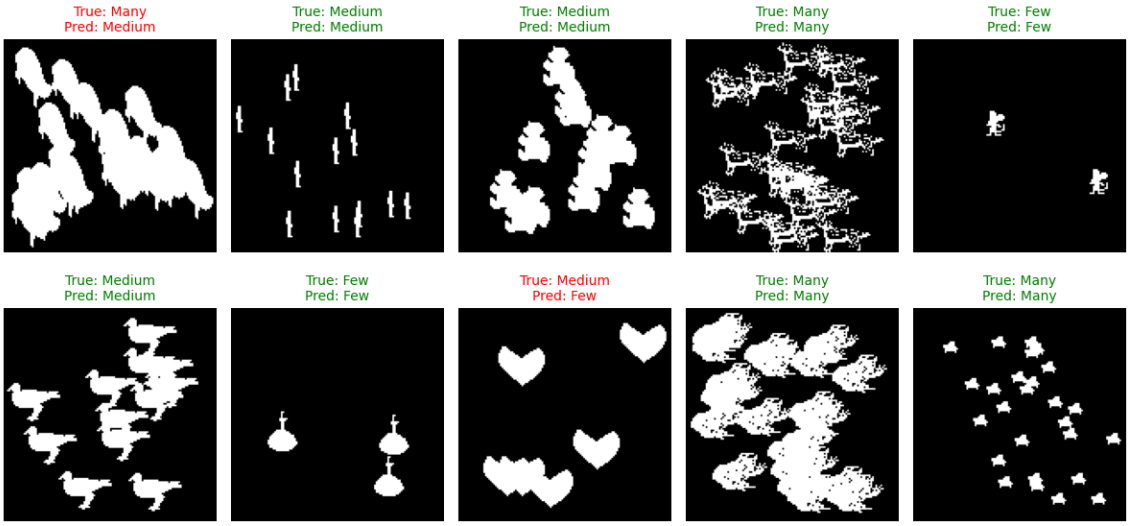

Figure 5.39: Sample Predictions from Fine-Tuned CNN (Dot → Silhouette). Correct predictions are highlighted in green and misclassifications in red. The CNN model struggled more with category separation compared to the Transformer-enhanced model.

CNN + Transformer Model

Figure 5.40 shows the training and validation curves for the fine-tuned CNN + Transformer. The model exhibited smooth convergence with lower validation loss and consistently higher accuracy than the CNN baseline.

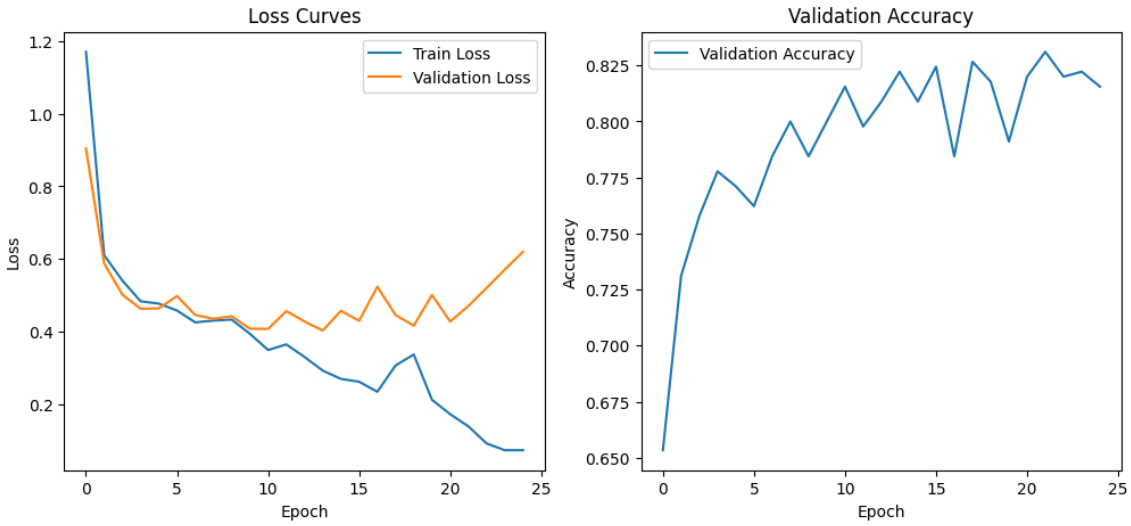

Figure 5.40: Training Metrics for Fine-Tuned CNN + Transformer (Dot → Silhouette). Left: Training vs Validation Loss. Right: Validation Accuracy over epochs. The hybrid model demonstrated smooth convergence and higher final accuracy compared to CNN alone.

The confusion matrix (Figure 5.41) reveals balanced performance across all categories, with reduced confusion in the "Medium" class relative to the CNN-only

model.

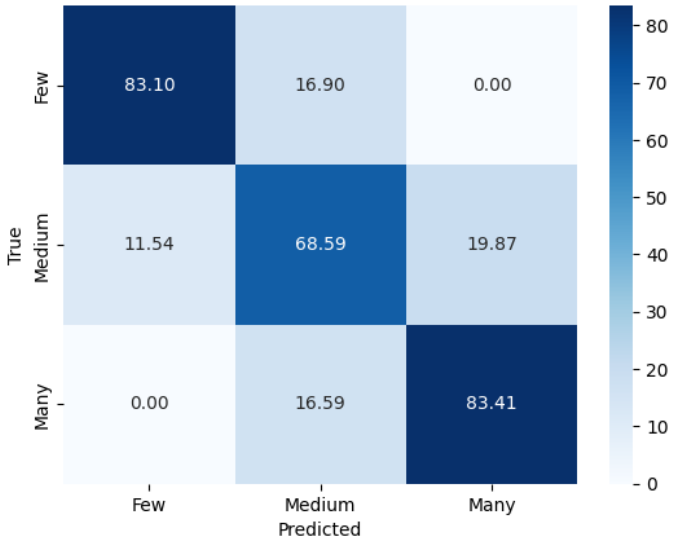

Figure 5.41: Confusion Matrix for Fine-Tuned CNN + Transformer (Dot → Silhouette). The model achieved better balance across "Few," "Medium," and "Many" classes compared to the pure CNN, although some confusion between "Medium" and "Many" remained.

Sample predictions (Figure 5.42) further illustrate the model's improved abstraction capabilities. Most classifications were correct and class boundaries were more distinct compared to the CNN model.

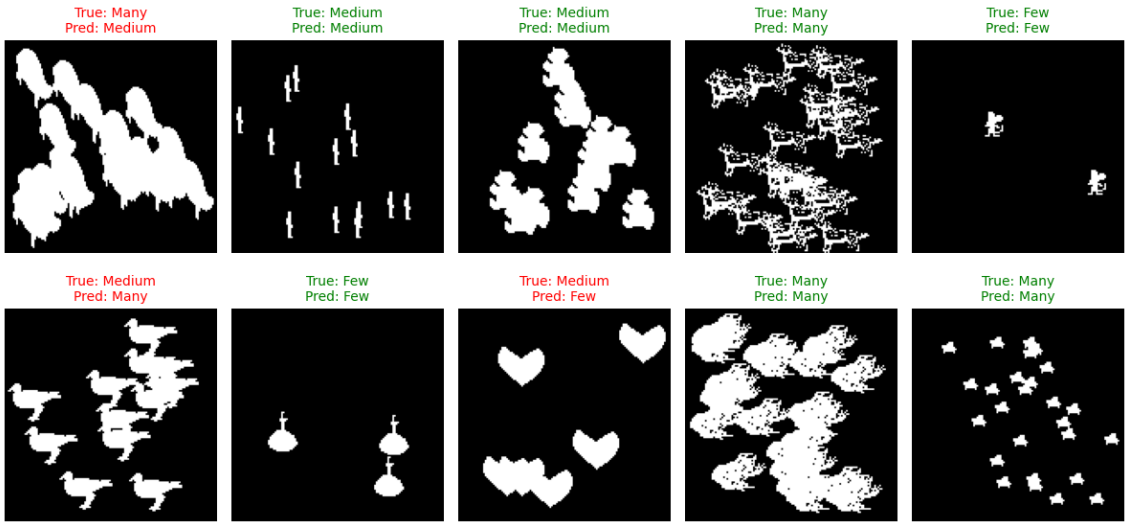

Figure 5.42: Sample Predictions from Fine-Tuned CNN + Transformer (Dot → Silhouette). Correct predictions are shown in green, misclassifications in red. The model displayed improved abstraction and category separation compared to the fine-tuned CNN model.

5.0.13 Evaluation of the Pixel-Ratio Hypothesis in Numerosity-Based Categorization

To evaluate whether models estimate white pixel density rather than abstract numerosity, a synthetic Pixel-Ratio dataset was constructed for controlled experiments. This section reports cross-dataset evaluations and pixel distribution analyses to probe this hypothesis.

Cross-Dataset Evaluation

We first trained both the CNN and CNN + Transformer models on the pixel-ratio-controlled dataset and evaluated their generalization to the dot and silhouette datasets. The results are presented in Table 5.21. While models achieved perfect accuracy on the pixel-ratio dataset, they failed to generalize, classifying nearly all real images as "Few." This suggests an overreliance on total white pixel count during training.

Conversely, Table 5.22 shows the performance when models trained on real datasets (Dot, Silhouette) were tested on the pixel-ratio dataset. These models achieved moderate accuracy, especially the silhouette-trained CNN, which generalized better than others. This asymmetry highlights that real-dataset-trained models internalize abstract numerosity to a limited degree, unlike pixel-trained models which fail completely under natural variations.

Table 5.21: Accuracy When Trained on Pixel-Ratio Dataset and Tested on Various Domains.

Test Dataset	**Model**	**Accuracy (%)**	**Observation**
Pixel-Ratio	CNN	100.00	Perfect fit
Pixel-Ratio	CNN + Transformer	100.00	Perfect fit
Dot	CNN	16.27	All samples classified as "Few"
Dot	CNN + Transformer	16.27	All samples classified as "Few"
Silhouette	CNN	15.78	Majority predicted as "Few"
Silhouette	CNN + Transformer	18.89	Slight increase in "Medium" predictions

Table 5.22: Accuracy When Tested on Pixel-Ratio Dataset with Dot/Silhouette-Trained Models.

Train Dataset	Model	Accuracy (%)	Observation
Dot	CNN	53.33	Moderate generalization
Dot	CNN + Transformer	56.11	Moderate generalization
Silhouette	CNN	66.11	Stronger generalization
Silhouette	CNN + Transformer	56.11	Moderate generalization

White Pixel Ratio Distributions

To further analyze this behavior, histograms of white pixel ratios were plotted for both the Dot and Silhouette datasets. Figure 5.43 shows that the "Few," "Medium," and "Many" classes have considerable overlap in pixel density, challenging any pixel-threshold-based classification strategy.

In Figure 5.44, this overlap is even more pronounced in the silhouette dataset, reinforcing the need for abstraction beyond simple pixel-level heuristics.

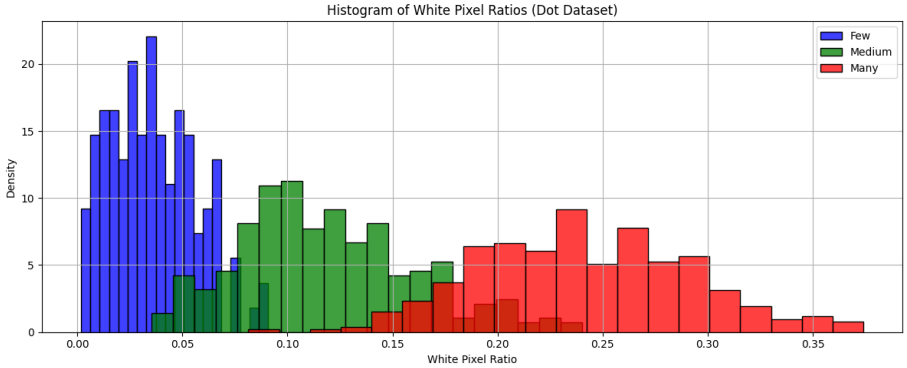

Figure 5.43: Histogram of White Pixel Ratios in the Dot Dataset. The "Few", "Medium" and "Many" classes show significant overlap in pixel density.

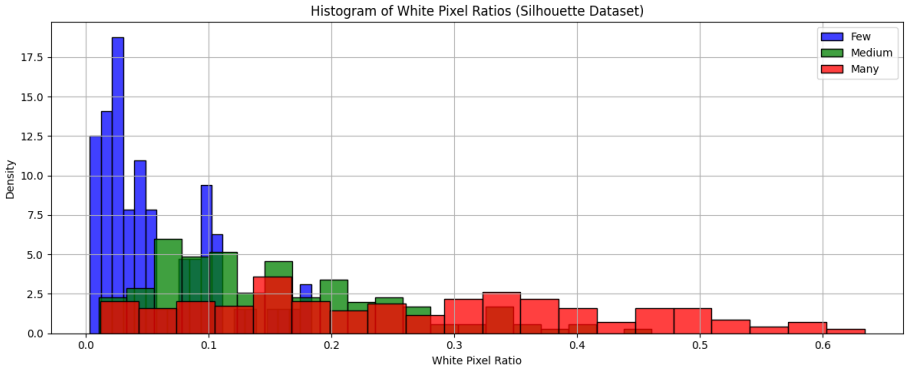

Figure 5.44: Histogram of White Pixel Ratios in the Silhouette Dataset. Overlaps between classes are even more pronounced than in the Dot Dataset.

Summary

Figure 5.43 and Figure 5.44 reveal that real datasets exhibit substantial pixel ratio overlap between classes. The results in Table 5.21 confirm that models trained only on pixel thresholds fail to generalize abstract numerosity. Meanwhile, Table 5.22 shows that real-dataset-trained models learn some abstraction, albeit limited. These observations support the conclusion that pixel coverage alone is not sufficient for genuine numerosity learning, a topic further analyzed in the Discussion chapter.

Chapter 6

Discussion

6.1 Overview of Key Findings

This thesis investigated the ability of neural networks to categorize abstract numerosity, "few", "medium" and "many" across varied visual domains. Experiments were conducted using synthetic dot patterns, silhouette compositions, pixel-ratio images and controlled dataset variations. Two model types were employed: a baseline CNN and a CNN+Transformer hybrid.

Across Experiments 1–7, both models achieved strong domain-specific performance, with the CNN+Transformer consistently outperforming the CNN in generalization scenarios. Fine-tuning experiments further revealed that cross-domain adaptation is feasible, albeit with moderate performance drops. Finally, the abstraction tests using pixel-ratio datasets provided conclusive evidence regarding the models' capacity to move beyond low-level pixel statistics.

6.2 Generalization Trends

Occlusion and Shape Variations

Both models generalized well to occlusion and shape transformations. Accuracy on occluded dot patterns exceeded 90%, demonstrating resilience to partial visibility. Similarly, replacing dots with varied geometric shapes had minimal impact on numerosity classification. This performance aligns with previous findings that neural networks can develop numerosity-sensitive mechanisms despite surface-level visual variation [19]

Clustered Variant: Underestimation Bias

A notable exception was observed in the clustered variant, where both CNN and CNN+Transformer models exhibited significant accuracy drops (CNN: 23.6%, Transformer: 36.27%). Confusion matrices revealed a consistent underestimation trend: "Many" instances were frequently misclassified as "Medium" or "Few".

This underestimation bias aligns with human perceptual limitations, where dense object groupings are often perceived as fewer distinct entities.

CNN: Struggled more severely due to its reliance on local receptive fields, causing tightly packed dots to merge in convolutional feature maps.

CNN+Transformer: Performed better, leveraging global self-attention for improved spatial reasoning; however, its gains were still constrained by the CNN backbone.

These results suggest that both models rely partially on density-sensitive visual cues but lack true spatial individuation mechanisms under clustering. Similar clustering induced underestimation was previously observed in both human cognition and CNNs trained on visual tasks, where where density overwhelmed individuation [21].

Future work should consider datasets explicitly incorporating dense clusters and architectures capable of object-level counting.

6.3 Abstraction Versus Surface Cue Reliance

To probe whether the models learned abstract numerosity or merely low-level pixel-based heuristics, experiments using a synthetic pixel-ratio dataset were conducted.

Models Trained on Pixel-Ratio

Models trained exclusively on pixel-ratio images failed to generalize to dot or silhouette datasets. Nearly all test samples were classified as "Few", regardless of actual numerosity.

Explanation: Real-world-like datasets exhibited substantially different white pixel distributions:

- "Many" class in dot images rarely exceeded 30% white pixels.

- Silhouette classes had heavy overlap across pixel ratios.

This mismatch invalidated pixel-threshold-based predictions, highlighting that pixel coverage alone is an unreliable numerosity proxy.

This supports the position in [21] that deep networks often regress to low-level texture statistics when abstraction is not explicitly encouraged.

Reverse Direction: Dot/Silhouette → Pixel-Ratio

Interestingly, models trained on dot and silhouette data generalized moderately well when tested on the pixel-ratio dataset (CNN achieving 53–66% accuracy). This suggests that such models learned a more abstract representation of numerosity, independent of white pixel density alone.

6.4 Semantic Mismatch in Pixel-Ratio Datasets

Further analysis revealed a crucial finding: even when pixel-ratio images were densely filled (e.g., a vertical bar covering 80% of pixels), models trained on dots or silhouettes often classified these images as "Few". This matches human intuition, a single large vertical bar is perceived as one object, not "many".

Thus, the classification mismatch was not a model error but a consequence of semantic misalignment between white pixel ratios and perceived numerosity. Models trained on object-based numerosity appeared to perform genuine object individuation, resisting misleading pixel-density cues.

The confusion matrices corroborate this:

- Pixel-ratio-trained models systematically misclassified dots and silhouettes as "Few".

- Dot- and silhouette-trained models partially succeeded in classifying pixel-ratio images, especially where patterns were spatially diversified (e.g., checkerbox).

This asymmetry strengthens the conclusion that neural models can learn perceptual numerosity abstraction, rather than simply encoding low-level pixel statistics.

6.5 Model Comparisons

Across most experiments, the CNN+Transformer hybrid model achieved superior performance compared to the CNN baseline. Global attention mechanisms enabled better handling of occlusion, domain transfer and abstract generalization.

However, both models exhibited similar cognitive biases, particularly under dense clustering, indicating that global attention alone is insufficient when the underlying feature extractor lacks individuation capabilities. Thus, hybrid architectures benefit generalization but do not eliminate fundamental perceptual limitations.

Recent work using neuro-symbolic modules to address similar limitations in subitizing suggests potential paths forward [1].

6.6 Cognitive Parallels and Human-Like Biases

The observed underestimation trends and generalization patterns suggest notable parallels to human numerosity perception.

Humans similarly:

- Struggle to individuate densely packed objects.
- Generalize approximate numerosity across varied layouts and shapes.
- Show resilience to occlusion while maintaining cardinality judgments.

These cognitive similarities imply that convolutional and attention-based models function as perceptual approximators rather than explicit counters, extracting distributed quantity features rather than itemized object counts.

This perceptual approach to quantity echoes the approximate number system proposed in neuroscience [12, 13] and modeled in abstract learning systems [19].

6.7 Context-Dependent Perception of Numerosity

In human cognition, numerosity perception is inherently context-dependent. The same object count can be perceived differently based on environmental scale and expectation, for instance, 15 people appear “many” in a small room but “few” in a large stadium.

This relativity highlights a challenge for computational models trained on fixed-scale, context-free datasets.

Interestingly, the CNN and CNN+Transformer models mirrored this human heuristic. In the clustered variant experiments, despite high object counts, dense spatial arrangements led to systematic underestimation, consistent with human tendencies to group proximate elements perceptually.

Furthermore, the failure of pixel-ratio-trained models to generalize contrasted with moderate success in dot/silhouette-trained models, this reinforces the notion that surface features like pixel density are insufficient for genuine numerosity abstraction.

Future research should explore incorporating explicit context awareness, scene-level spatial priors and hierarchical reasoning modules to develop models capable of human-aligned numerosity abstraction.

6.8 Summary

The results of this thesis collectively demonstrate that while modern neural networks can approximate human-like numerosity perception under controlled conditions, important challenges remain in achieving robust, context-independent abstraction. Observed biases, particularly under clustering and density manipulation highlight the need for future models to integrate not only global spatial reasoning, but also mechanisms for object individuation and flexible context interpretation. Ultimately, advancing numerosity modeling will require bridging the gap between perceptual cues and conceptual understanding, mirroring the intricate interplay observed in human cognition.

Chapter 7

Conclusion and Future Work

This thesis explored the capacity of neural networks to categorize numerosity into abstract classes such as "few," "medium," and "many." Through a comprehensive series of experiments across synthetic dot patterns, silhouette compositions and pixel-ratio abstractions, the study systematically evaluated generalization, cross-domain transfer and potential perceptual biases in neural quantity estimation.

The results demonstrate that neural models can, to a significant extent, abstract numerosity beyond surface features such as object identity and spatial arrangement. Both the CNN and CNN+Transformer architectures performed well on in-domain test sets and exhibited resilience to visual perturbations such as shape changes and occlusion. However, notable limitations emerged when models were tested on densely clustered inputs or when trained exclusively on context-free pixel-ratio distributions.

The pixel-ratio experiments were particularly revealing. Models trained solely on white pixel density failed to generalize to object-based datasets, whereas models trained on dots and silhouettes generalized moderately well to pixel-ratio tasks. This asymmetry strongly suggests that the learned representations are based on abstract numerosity, rather than simple pixel statistics.

The consistent underestimation bias observed in clustered images further parallels human perceptual heuristics, where spatial compactness can lead to reduced perceived quantity. Such behaviors highlight the emergence of cognitive-like approximations in neural models, offering valuable insights into both artificial and biological numerosity perception.

7.1 Limitations

Despite the contributions, several limitations of this study should be acknowledged:

- **Synthetic Datasets:** All experiments were conducted on synthetic data. Real-world scenes, with their inherent noise, texture variability and complex occlusions, may present additional challenges.
- **Input Resolution:** Models were trained on 128×128 grayscale images. Larger and more varied input resolutions could impact generalization and abstraction performance.
- **Model Complexity:** The Transformer architecture employed was relatively shallow and the potential of deep self-attention mechanisms may not have been fully exploited.

Furthermore, the experiments did not explicitly incorporate *contextual relativity*, an essential aspect of human numerosity perception. In real-world settings, the perceived quantity depends not only on absolute count but also on spatial scale and semantic context, for instance, 15 people may feel "many" in a small room but "few" in a stadium. Capturing such flexible, context-dependent numerosity remains an open challenge for future modeling efforts.

7.2 Future Work

Building on the findings of this thesis, several promising directions are identified for future research:

- **Data Complexity and Augmentation:** Extend datasets to include denser clustering, more irregular spatial patterns, real-world textures and occlusions. This would better simulate naturalistic scenarios and challenge models to develop stronger abstraction abilities.
- **Cross-Modality Expansion:** Explore numerosity categorization beyond visual inputs. New modalities could include:

  - *Temporal sequences:* Classify sequences based on the number of peaks or spikes.
  - *Graphs:* Treat nodes as items and classify based on node counts or connectivity patterns.
  - *Audio signals:* Estimate quantity from patterns like repeated sounds or frequency bursts.

  Such expansions would test whether learned representations of numerosity can generalize beyond visual spatial layouts to abstract structural and temporal domains.

- **Attention Mechanism Analysis:** Visualize attention maps (especially in Transformer-based models) to understand which spatial regions or features the models rely on when estimating numerosity.

- **Ordinal-Aware Learning Objectives:** Replace standard categorical cross-entropy loss with ordinal-aware loss functions. Since numerosity classes (Few, Medium, Many) are naturally ordered, ordinal loss could enhance classification fidelity and model calibration.

- **Multi-Scale and Context-Aware Architectures:** Investigate hierarchical or pyramid-based models that process information at multiple spatial scales. Incorporating broader scene context may help mitigate errors under clustering or scale variations.

- **Hybrid Symbolic-Perceptual Reasoning:** Develop hybrid models that combine perceptual feature extraction with symbolic counting modules. For example, explicit object individuation heads could be integrated to bridge the gap between density perception and itemized numerosity estimation.

- **Curriculum Learning Strategies:** Implement progressive learning setups where models first train on simpler dot patterns, then progressively move to silhouettes, pixel-ratio images, graphs, sequences, and finally real-world data. This would mirror developmental trajectories observed in human numerosity acquisition.

These avenues aim not only to improve model robustness but also to push neural networks toward a more human-like, context-aware and modality-agnostic understanding of quantity.

7.3 Final Remarks

This research contributes to the growing body of work at the intersection of neural learning and cognitive abstraction. It demonstrates that, with careful experimental design and model structuring, neural networks can begin to approximate flexible, human-like representations of quantity. These findings offer a foundation for the development of more generalizable, perceptually grounded AI systems, capable of moving beyond fixed surface features toward deeper cognitive abstraction.

Bibliography

[1] Md. Harunur Rashid Alam, Tengyu Ma, and Yu-Xiong Wang. "Towards Generalization in Subitizing with Neuro-Symbolic Loss using Holographic Reduced Representations". In: *arXiv preprint arXiv:2401.00003* (2024). URL: https://openreview.net/pdf?id=AOAP8sLYdt.

[2] Stanislas Dehaene. *The Number Sense: How the Mind Creates Mathematics*. Oxford University Press, 1997.

[3] Lijia Deng et al. "Deep Learning for Crowd Counting: A Survey". In: *CAAI Transactions on Intelligence Technology* 9 (2023), pp. 1043–1077. DOI: 10.1049/cit2.12241.

[4] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *International Conference on Machine Learning (ICML)*. 2017, pp. 1126–1135. URL: https://arxiv.org/abs/1703.03400.

[6] Ben M Harvey and Serge O Dumoulin. "A network of topographic numerosity maps in human association cortex". In: *Nature Human Behaviour* 1 (2017), p. 491. DOI: 10.1038/s41562-016-0036.

[7] Ben M Harvey et al. "Topographic representation of numerosity in the human parietal cortex". In: *Science* 341.6150 (2013), pp. 1123–1126. DOI: 10.1126/science.1239052.

[8] Janusz Kacprzyk, Mario Fedrizzi, and Hannu Nurmi. *Fuzzy Logic with Linguistic Quantifiers in Group Decision Making*. Ed. by Lotfi A. Yager Ronald R.and Zadeh. Springer US, 1992, pp. 263–280. DOI: 10.1007/978-1-4615-3640-6_13.

[9] Longin Jan Latecki and Richard Ralph. *MPEG-7 CE-Shape-1 Part B Dataset*. Accessed: 2025-03. 2001. URL: https://dabi.temple.edu/external/shape/MPEG7/dataset.html.

[10] Muzammal Naseer et al. "Intriguing Properties of Vision Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 11986–11995. URL: https://arxiv.org/abs/2105.10497.

[11] Khaled Nasr, Prem Viswanathan, and Andreas Nieder. "Number detectors spontaneously emerge in a deep neural network designed for visual object recognition". In: *Science Advances* 5.5 (2019), eaav7903. DOI: 10.1126/sciadv.aav7903.

[12] Andreas Nieder. "Counting on neurons: The neurobiology of numerical competence". In: *Nature Reviews Neuroscience* 6.3 (2005), pp. 177–190. DOI: 10.1038/nrn1626.

[13] Andreas Nieder and Stanislas Dehaene. "Representation of number in the brain". In: *Annual Review of Neuroscience* 32 (2009), pp. 185–208. DOI: 10.1146/annurev.neuro.051508.135550.

[14] George Papamakarios. "Neural Density Estimation and Likelihood-free Inference". In: *arXiv* (2019). URL: https://arxiv.org/abs/1910.13233.

[15] Manuela Piazza and Veronique Izard. "Developmental trajectories of number acuity in humans". In: *Trends in Cognitive Sciences* 14.6 (2010), pp. 289–296. DOI: 10.1016/j.cognition.2010.03.012.

[16] Rui Jiang Qiao Liu Jiaze Xu and Wing Hung Wong. "Density estimation using deep generative neural networks". In: *Proceedings of the National Academy of Sciences* (2021). DOI: 10.1073/pnas.2101344118.

[17] Maithra Raghu et al. "Do Vision Transformers See Like Convolutional Neural Networks?" In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. 2022, pp. 12116–12128. DOI: 10.48550/arXiv.2108.08810.

[18] Vishwanath A. Sindagi and Vishal M. Patel. "A survey of recent advances in CNN-based single image crowd counting and density estimation". In: *Pattern Recognition Letters* 107 (2018), pp. 3–16. DOI: 10.1016/j.patrec.2017.07.007.

[19] Ivilin Stoianov and Marco Zorzi. "Emergence of a "visual number sense" in hierarchical generative models". In: *Nature Neuroscience* 15.2 (2012), pp. 194–196. DOI: 10.1038/nn.2996.

[20] Taylor W. Webb, Ishan Sinha, and Jonathan D. Cohen. "Emergent Symbols through Binding in External Memory". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://arxiv.org/abs/2012.14601.

[21] Jiayu Wu, Tengyu Ma, and Yu-Xiong Wang. "Cognitive deficit of deep learning in numerosity". In: *arXiv preprint arXiv:1905.02262* (2019). DOI: 10.1609/aaai.v33i01.33011303.

[22] Xiangliang Yuan et al. "The neural correlates of individual differences in numerosity perception". In: *Cerebral Cortex* 33.12 (2023), pp. 7698–7711. DOI: 10.1016/j.isci.2023.107392.