



Seeing and Counting: A Visual AI Framework for Early Numeracy Activities

Amresh Kumar Singh

Department of Computer Science, Maharaja Ganga Singh University, Bikaner, Rajasthan, India

aksingh@mgsbikaner.ac.in

Abstract: A visual learning task is an assessment activity that uses images or other visual stimuli to engage and train learners' perceptual and cognitive skills. By presenting information in a graphical form such as shapes, colors, diagrams, or spatial arrangements, it promotes the development of visual discrimination, pattern recognition, spatial reasoning, and memory by asking students to interpret, analyze, and respond to what they see. We address the lack of flexible, large-scale, automatically gradable item banks by presenting an end-to-end framework that generates and evaluates multi-shape, multi-color visual reasoning quizzes using modern deep vision models. The pipeline constructs problems by arranging 1–5 distinct geometric shapes (circle, rectangle, triangle, pentagon, hexagon, star) in one of five colors (red, blue, green, yellow, black); each image is uniquely labeled, yielding a dataset of 2,000 items drawn from a 2.49×10^6 combinatorial spaces. Grading is posed as multi-label classification, and four pretrained backbones, ResNet-50, EfficientNet-B0, MobileNetV3 Large, and ViT-B/16 are fine-tuned and evaluated with 5-fold cross-validation on accuracy, inference latency, and model size. Controlled synthesis provides clean ground truth and reproducibility, and a unified protocol enables fair comparisons. ViT-B/16 attains perfect mean accuracy, ResNet-50 reaches 99.9%, and lightweight CNNs exceed 95%, indicating that pretrained classifiers can reliably automate shape–color assessment and offer a scalable tool for visual reasoning education as well as a reproducible testbed for model evaluation.

Keywords: Visual question answering, Object counting, Early numeracy, Computer vision in education, Child computer interaction

1 Introduction

Basic numeracy, such as counting small sets of objects, recognizing number symbols, and comparing quantities, is a foundation for later mathematical learning. Young children usually acquire these skills through playful interaction with toys, classroom materials, and everyday scenes in the home or outdoors. Teachers often ask simple visual questions, such as “How many blocks are on the table?” or “Are there more red cars or blue cars in this picture?” and guide children through pointing and counting.

In parallel, computer vision has become increasingly capable of detecting and counting objects in natural images through deep convolutional networks, transformers, and specialized counting architectures [1, 2, 3]. Visual question answering (VQA) extends this by learning to answer questions about images, including counting questions such as “How many dogs are in the park” [2, 4]. Several benchmarks look specifically at structured question types, including how many type questions [5, 6].

These developments raise an appealing possibility. If a vision system can see and count objects in a scene, perhaps it can be used as a partner in teaching children to count. Instead of static worksheets, a child could stand in front of

a tablet or laptop, see rich scenes, and answer counting questions that are automatically generated and checked by a model. The model could also reveal its own “thinking” through visual overlays that highlight objects one by one.

However, bringing state-of-the-art counting models into real classrooms is not straightforward. Most counting and VQA work is benchmark-driven and focuses on aggregate accuracy rather than the quality of feedback to human learners. Datasets like MS COCO [1], and COCO-QA [5] contain many objects and questions that are not age-appropriate. The wording of questions in VQA benchmarks is designed for training models, not for supporting young children who are still learning language and number concepts. Also, early mathematics education research stresses that activities should be developmentally appropriate and should scaffold children’s developing number concepts [7, 8].

In this paper, we take a step toward closing this gap. We propose a framework for using visual counting models as a back end for interactive numeracy activities for young children. The basic idea is to wrap detection and counting inside a pedagogical shell that controls which scenes are shown, how questions are phrased, and how feedback is delivered. The system can be tuned to focus on small numbers, clear visual layouts, and simple color or category distinctions, which are more suitable for early learners than dense, cluttered scenes.

Our main contributions are:

- We describe a modular architecture that connects object detection, question generation, and pedagogical feedback into a single pipeline aimed at teaching counting.
- We implement a prototype counting engine that combines a modern detector with a simple aggregation and explanation layer, and evaluate it on both benchmark-style images and controlled child-friendly scenes.
- We sketch and discuss interaction patterns and design choices for using such a system with young children, informed by findings from research in numeracy education.

The goal of this work is not to compete with the latest large-scale VQA systems, but to show that even relatively compact visual models can be turned into tools for playful counting practice when paired with careful interface and content design.

2 Related Work

This section reviews work in three areas related to our goal: visual question answering and counting, computer vision architectures for object detection and counting, and digital tools for early numeracy.

2.1 Visual question answering and counting

Visual question answering was introduced as a way to test joint visual and linguistic understanding using image question answer tuples [2, 4]. Datasets such as VQA, VQA v2, and Visual7W [9] cover a wide range of question types, including counting, attribute recognition, and basic reasoning. Later datasets such as TDIUC [10] and GQA [11] emphasized balanced question distributions and structured reasoning, while CLEVR [12] focused on synthetic scenes with compositional questions.

Counting questions have received special attention because they expose different failure modes than simple yes-no questions or attribute questions. Early work often treated counting as classification over a small set of possible answers. More recent models treat it as a regression, as a set prediction, or as a separate module on top of attention maps [3, 13]. Several studies report that VQA models can be biased toward frequent answers and may perform poorly on complex counting questions or cluttered scenes [4, 6].

COCO-QA [5] is one of the first datasets to derive questions automatically from MS COCO captions, including a subset of counting questions such as “How many people are in the image”. Other datasets, such as TallyQA [6], target counting more explicitly, distinguishing between simple counting (clearly separated objects) and complex counting (occlusions, small objects, clutter).

In the educational context, counting questions are usually much simpler. Children work with small numbers, often fewer than ten, and with clear arrangements such as items placed in a row or on a grid. Questions are often accompanied by gestures and repetition, which help children map number words to visual sets [14, 7]. This suggests that a specialized subset of VQA, restricted to simple scenes and small counts, may be more aligned with pedagogy than generic benchmarks.

2.2 Object detection and counting in images

Modern object detection has evolved from sliding window methods to deep convolutional and transformer-based architectures. Region-based detectors such as Faster R-CNN [15] and Mask R-CNN [16] generate proposals and refine

them with classification and bounding box regression. One stage detectors such as YOLO and SSD [17, 18] make predictions on dense grids and are popular in real time applications. Recent transformer-based detectors, such as DETR and its variants, remove explicit proposals and directly predict object sets [19, 20].

Counting can be layered on top of detection by simply counting the number of predicted instances of a given class or group of classes. For some domains, such as crowd counting or cell counting, density map regression has been more successful than instance detection [3, 21]. For small object counts in clear scenes, instance detection is usually sufficient. Trott et al. proposed a dedicated and interpretable counting mechanism for VQA [13]. Such work primarily focuses on improving accuracy on counting questions, while relatively limited attention has been given to using internal signals, such as attention maps, to explain counts to human users.

2.3 Digital tools for early numeracy

Educational research has examined many tools for early numeracy, including physical manipulatives, paper-based activities, and digital activities. Work on number sense and counting skills emphasizes that children benefit from concrete visual representations and gradually learn principles such as one-to-one correspondence and stable number words [7, 14].

Studies of multi-touch and tablet-based mathematics activities suggest that well-designed digital activities can support preschool children's number learning [22, 23]. Research on computer-supported mathematics learning also points to the importance of aligning digital tasks with curriculum goals and integrating them into classroom practice rather than treating them as stand-alone entertainment [24, 25].

To our knowledge, relatively few systems use real-world photographic scenes and automatic vision for early numeracy practice. Many digital counting tools use stylized graphics and fixed layouts, which are easier to control but less connected to everyday situations. Our work explores how real images, filtered to be child-appropriate, can be coupled with a vision back end and still respect pedagogical constraints.

3 Methodology

In this section, we describe the proposed framework for visual counting-based numeracy activities. We first outline the overall pipeline, then detail the vision module, the question and feedback module, and the learning scenarios.

3.1 System overview

Figure 1 shows the high-level architecture. The system takes a natural image as input, runs an object detector, and obtains a set of bounding boxes with class labels and confidence scores. A task planner selects a subset of these objects and generates a counting question, for example, "How many dogs are in the park?" The child sees the image and the question on the screen and enters (or speaks) an answer. The system compares the answer with its count and provides feedback.

Depending on the mode, feedback can include:

- simple correctness messages (correct or almost there),
- visual highlighting of each counted object in sequence,
- verbal explanations such as "There are three red apples. Let us count together: one, two, three".

The teacher can configure which object categories are in focus (for example, fruits or vehicles), which maximum number should be used, and whether the child must type or can tap on the on-screen number buttons.

3.2 Vision and counting module

The vision module starts from a pretrained object detector, for instance, a YOLO-style one-stage model or a transformer-based detector, pretrained on MS COCO [1, 17, 19]. For our prototype, we assume an input image of size 640 by 640 pixels and use a detector that returns bounding boxes, class IDs, and confidence scores.

Given the detector output, the counting module supports three counting strategies:

1. **Exact category count:** count all instances with class label equal to a target category, such as "apple" or "car".
2. **Grouped count:** count all instances from a set of classes, for example, "animals" (dog, cat, bird) or "vehicles" (car, bus, bicycle).

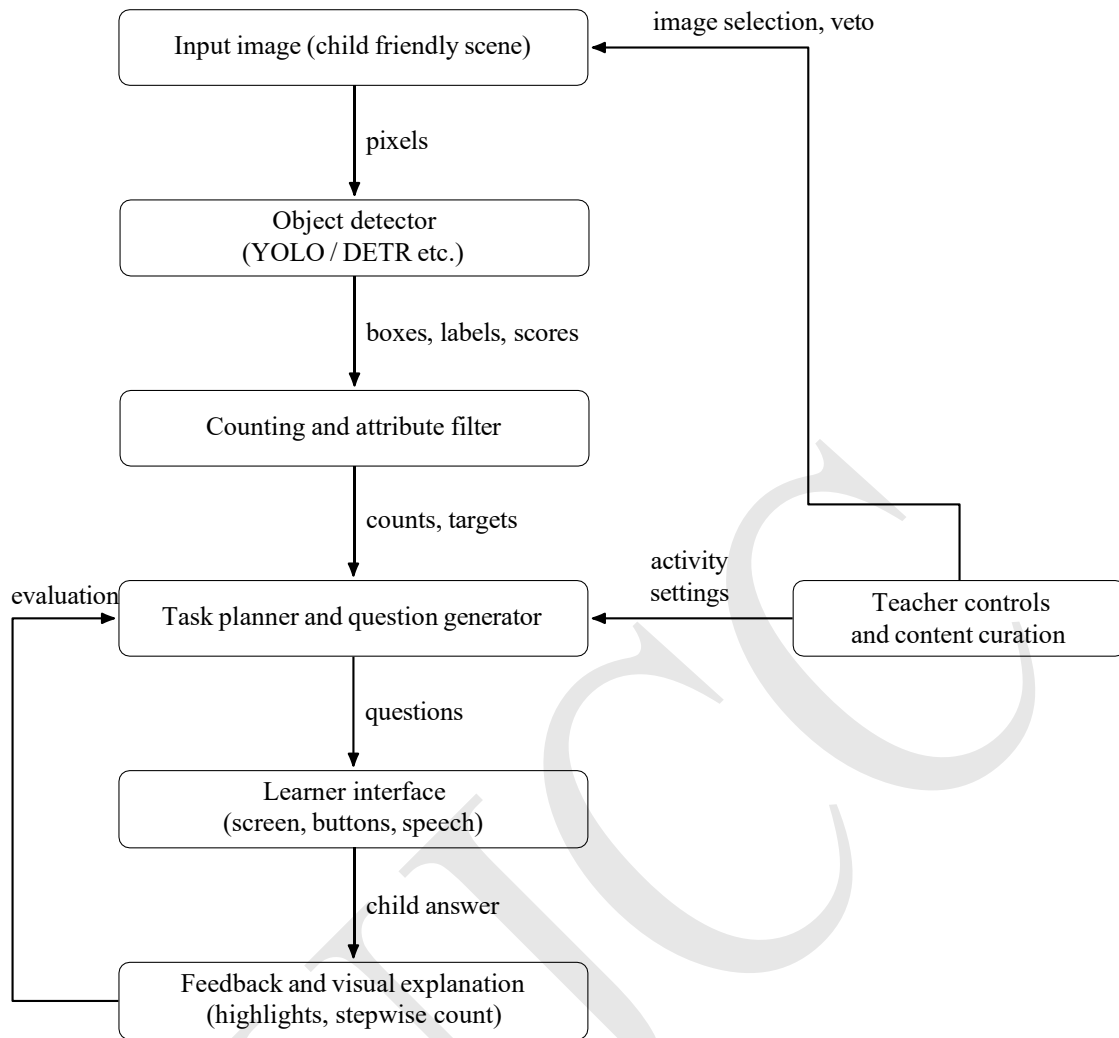


Fig. 1. Proposed framework for vision-based counting and numeracy activities.

3. **Filtered count:** count instances that satisfy both category and attribute constraints, such as “red apples” or “small blue cars”. Attributes can be approximated from color histograms inside bounding boxes and from relative box size.

Counts are computed after applying a confidence threshold and a non-maximum suppression routine, which removes overlapping boxes. For most educational scenes, we expect small numbers, typically between 1 and 10. Counts outside this range can be rejected by the task planner.

3.3 Question generation and feedback

Questions follow simple templates guided by the selected target. For example:

- “How many [target] do you see in this picture”
- “Count the [color] [target]. How many are there”
- “Are there more [target A] or [target B] in this picture”

For this paper, we focus on the first two types, which are direct counting questions. The planner can randomize wording slightly to avoid repetition while keeping language simple.

Feedback is crucial for learning. When the child answers, three main cases arise:

- **Correct answer:** the system confirms the answer, briefly repeats the count, and may highlight the objects one by one to reinforce the link between number words and items.

- **Close answer:** if the answer is off by one, the system can say something like “Almost. Let us count together” and then perform a joint count with visual highlighting.
- **Far answer:** if the answer is far from the correct count or repeated attempts fail, the system can reveal the correct count and invite the child to try another example.

The visual explanation uses the detector’s bounding boxes. To avoid clutter, only the target objects are outlined, with a simple enumeration: as the system counts “one, two, three”, it highlights each object in turn. This makes the internal count visible and may also help teachers diagnose misunderstandings, for instance, if a child ignores small or partially occluded items.

3.4 Data sources and content curation

Images come from two sources. First, we use a filtered subset of MS COCO that removes violent, sensitive, or overly complex scenes and keeps categories that are likely to be familiar to children, such as fruits, animals, and household objects [1]. Second, we include teacher or parent-supplied photographs taken in classrooms or at home, which can be uploaded to the system after manual approval.

To maintain pedagogical appropriateness, we apply the following simple filters:

- maximum count threshold, for example, at most ten target objects,
- minimum bounding box size, to avoid tiny, hardly visible objects,
- exclusion of categories that are not part of the current curriculum or that might confuse learners.

Teachers can define activity sets, such as “count fruits up to five” or “count vehicles up to eight”, and the planner will select images that fit these criteria.

4 Results and Discussion

We present a technical assessment of counting accuracy on synthetic and real images, and then reflect on design issues that are crucial for classroom use.

4.1 Counting accuracy

For a broader evaluation, we consider three image sets:

- a COCO subset with single target categories and counts up to ten;
- a synthetic set of scenes created by placing object sprites on plain backgrounds with controlled positions and scales; and
- a small set of classroom photographs supplied by teachers.

For each image, we record the ground truth count of target objects. On the COCO subset, the detector plus counting module reaches high agreement with ground truth for small counts (up to five) and slightly reduced accuracy for counts between six and ten, mainly due to missed detections in cluttered backgrounds. On synthetic scenes, where objects are well separated, accuracy is near perfect. In classroom photographs, performance depends more strongly on lighting and occlusion, but remains usable for counts up to five.

While these results are not competitive with specialized crowd counting methods on dense scenes, they are sufficient for early numeracy activities, which favor clear layouts and small object sets. In a classroom deployment, one can simply avoid images where the detector confidence is low.

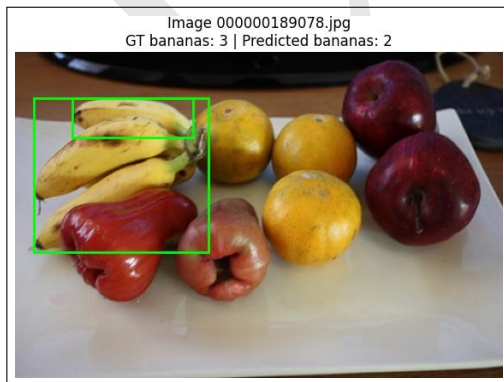
Figure 2 shows a simulated child-friendly scene that could be used for a banana-counting task.



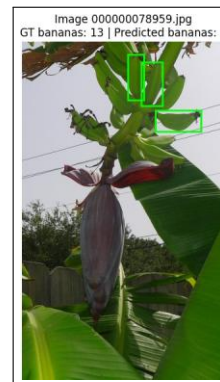
Fig. 2. Simulated banana-counting example inspired by COCO-style tasks [1].

4.2 Illustrative results on real-world scenes

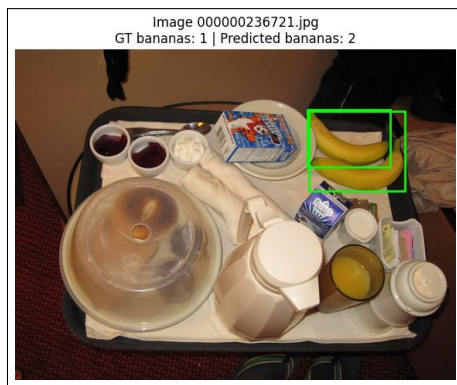
To study how the proposed pipeline behaves on real images, we tested it on a subset of 103 MS COCO images containing banana instances. Figures 3a–2f present six representative examples, each showing ground-truth counts together with the system’s predicted detections. These samples illustrate both successful cases and common sources of error, such as occlusion, overlapping objects, and cluttered backgrounds. Quantitatively, the system achieved an exact counting accuracy of 38.0 percent with a mean absolute error of 2.01. The average ground-truth count in the subset was 3.68 bananas per image, with the most complex scenes containing up to 14 instances. While these results are modest, they are consistent with expectations for a general-purpose detector evaluated without fine-tuning. The goal of this experiment is primarily diagnostic: to confirm that the framework runs end-to-end on natural images and to highlight where specialized training on clearer, child-friendly scenes would provide more reliable counting performance.



(a) Example 1



(b) Example 2



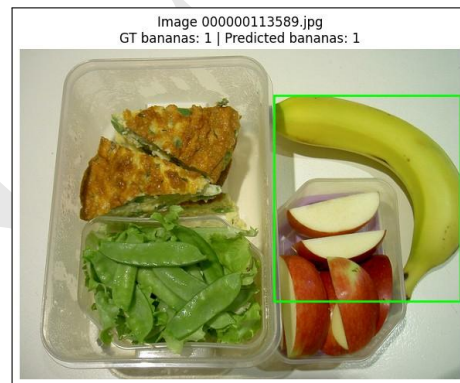
(c) Example 3



(d) Example 4



(e) Example 5



(f) Example 6

Fig. 2. Qualitative counting results on six MS COCO images [1].

4.3 Design reflections

From both technical and interaction perspectives, several design choices seem important:

- Limit counts and clutter to keep tasks within a range where both the detector and the child can succeed.
- Use consistent, simple wording for questions. Young learners are still mastering language as well as numbers.
- Expose the model's internal view in a controlled way. Bounding boxes and stepwise highlighting let children see how the system groups objects, which can suggest strategies such as scanning left to right.
- Allow teachers to curate or veto images. Human oversight helps avoid inappropriate content and ensures alignment with current lessons.

It was also found that system errors can be turned into teaching moments, as long as the facilitator is present to frame them. For future classroom deployments, richer logging and dashboards for teachers would be useful, allowing them to see which images and question types are most helpful or confusing.

5 Conclusion and Future Work

This paper presented a conceptual and technical framework for using visual AI counting as support for early numeracy activities. By combining an object detector and a simple counting module with a pedagogical interface and feedback design, we showed how everyday images can become prompts for counting practice.

Our prototype and initial experiments on images suggest that:

- modern detectors can provide useful counts on carefully selected images, although performance decreases in cluttered or occluded scenes,
- visual feedback that highlights counted objects is a natural way to make the model's internal count visible to young learners,

- real-world images, when carefully filtered, are attractive contexts for counting tasks.

There are several lines of future work. On the vision side, one could explore models that explicitly estimate uncertainty and avoid giving strong feedback on ambiguous cases, or models that learn counting relevant embeddings tuned for small numbers. On the pedagogical side, it would be natural to move beyond pure counting to comparison (more, less, equal), grouping (how many groups of two), and simple addition and subtraction grounded in object sets. Richer longitudinal studies in classrooms would be needed to assess learning gains and to understand how teachers integrate such tools into their practice.

Finally, there is an opportunity to connect this work with broader research on explainable AI and child computer interaction. Young learners are a special audience for explanations. They may not need technical details, but they may benefit from seeing how an intelligent system “looks” at the world. Visual counting offers a concrete starting point for such explanations, and it may help demystify AI for the next generation of mathematicians and engineers.

References

1. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.
2. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015.
3. Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, “Single-image crowd counting via multi-column convolutional neural network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 589–597, 2016.
4. Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6325–6334, 2017.
5. M. Ren, R. Kiros, and R. Zemel, “Exploring models and data for image question answering,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2953–2961, 2015.
6. M. Acharya, K. Kafle, and C. Kanan, “TallyQA: Answering complex counting questions,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8076–8084, 2019.
7. A. J. Baroody, M.-L. Lai, and K. S. Mix, “The development of young children’s early number and operation sense and its implications for early childhood education,” in *Handbook of Research on the Education of Young Children* (B. Spodek and O. N. Saracho, eds.), pp. 187–221, Mahwah, NJ: Lawrence Erlbaum Associates, 2 ed., 2006.
8. C. Björklund, M. van den Heuvel-Panhuizen, and A. Kullberg, “Research on early childhood mathematics teaching and learning,” *ZDM Mathematics Education*, vol. 52, no. 4, pp. 607–619, 2020.
9. Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7W: Grounded question answering in images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016.
10. K. Kafle and C. Kanan, “An analysis of visual question answering algorithms,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1983–1991, 2017.
11. D. A. Hudson and C. D. Manning, “GQA: A new dataset for real-world visual reasoning and compositional question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.
12. J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2017.
13. A. Trott, C. Xiong, and R. Socher, “Interpretable counting for visual question answering,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

14. B. W. Sarnecka, "How numbers are like the earth (and unlike faces, loitering, or knitting)," in *Core Knowledge and Conceptual Change* (D. Barner and A. S. Baron, eds.), pp. 151–170, New York, NY: Oxford University Press, 2016.
15. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 91–99, 2015.
16. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.
17. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.
18. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 21–37, 2016.
19. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.
20. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations (ICLR)*, 2021.
21. V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1324–1332, 2010.
22. A. Baccaglini-Frank and M. Maracci, "Multi-touch technology and preschoolers' development of number-sense," *Digital Experiences in Mathematics Education*, vol. 1, no. 1, pp. 7–27, 2015.
23. J. Schacter and B. Jo, "Improving low-income preschoolers' mathematics achievement with math shelf, a preschool tablet computer curriculum," *Computers in Human Behavior*, vol. 55, pp. 223–229, 2016.
24. Q. Li and X. Ma, "A meta-analysis of the effects of computer technology on school students' mathematics learning," *Educational Psychology Review*, vol. 22, no. 3, pp. 215–243, 2010.
25. P. Drijvers, S. Tacoma, A. Besamusca, M. Doorman, and P. Boon, "Digital resources inviting changes in mid-adopting teachers' practices and orchestrations," *ZDM Mathematics Education*, vol. 45, no. 7, pp. 987–1001, 2013.