THE MIXCOUNT DATASET: EXACT ANNOTATIONS AT SCALE FOR OPEN-VOCABULARY OBJECT COUNTING

Anonymous authors
Paper under double-blind review



Figure 1: **The MixCount Dataset.** Our dataset includes images of mixed objects in settings that are challenging for visual counting models, along with precise and rich ground truth annotations, text descriptions and visual exemplars.

ABSTRACT

Object counting is a foundational vision task with over a decade of dedicated research, yet state-of-the-art models still fail systematically when evaluated on real-world data with high total count, mixed objects, or a complex background. We show that this gap is strongly driven by limitations in existing training and evaluation data. Annotating real data is hardest exactly where it matters most, since a dense scene has to be labeled instance by instance, and existing datasets contain numerous errors. Synthetic alternatives remove the annotation cost but lack diversity and realism. We address this with MixCount, a dataset and benchmark for object counting whose annotations are exact by construction. An automatic pipeline synthesizes images, fine-grained textual descriptions and annotations at scale, eliminating the labeling errors that plague prior datasets. Evaluating state-of-the-art models on MixCount exposes severe degradation and provides insights into their sensitivity to data scale, exemplar quality, and object visibility or size. More importantly, augmenting their training with our synthesized data yields substantial gains on real-world benchmarks, reducing MAE by 20.14% on FSC-147 and up to 39.11% on PairTally. These results establish MixCount as both a benchmark and a training dataset for fine-grained counting, and demonstrate that exact annotations at scale help address a long-standing data bottleneck in counting models.

1 INTRODUCTION

Visual object counting has been a fundamental task in computer vision for several decades. Many real-world applications ranging from cell-counting (Phelan & Lawler, 1997; Xie et al., 2018), crowd-counting (Zhang et al., 2015), wildlife monitoring (Arteta et al., 2016), or industrial inspection (Khule et al., 2015) heavily rely on automated counting techniques to save significant time and enable further data analysis. In recent years, the capabilities of counting models have greatly expanded. General models that can count *anything* (Ranjan et al., 2021) have been proposed. They can be prompted to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

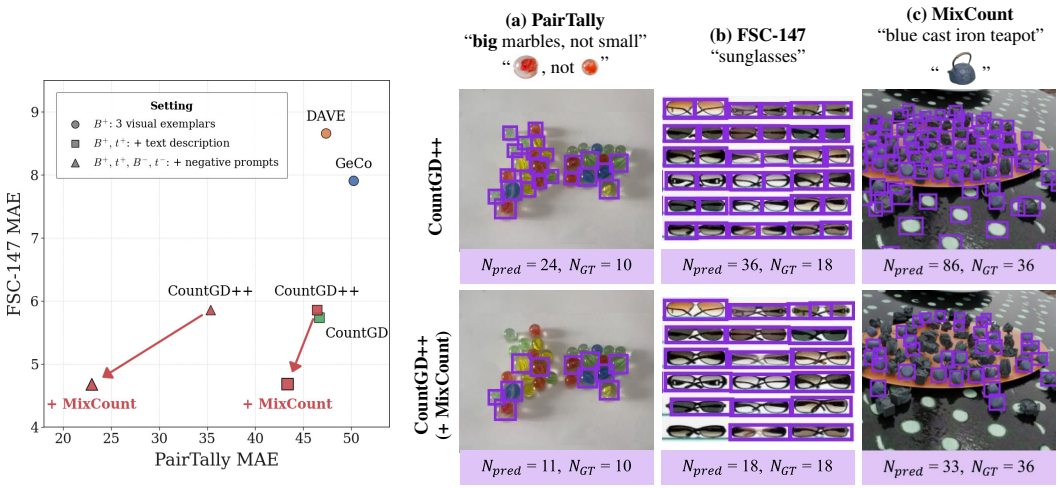


Figure 2: **Bridging the data gap.** Visual counting models often struggle to distinguish similar objects (a), fail to recognize self-similar components as a single entity (b), and are easily distracted by repetitive background patterns (c). We show that these failure modes can be substantially reduced with targeted training data, and design MixCount to bridge this gap, resulting in -20% error on recent benchmarks. Prompts are shown above the images.

count a specific class of objects via an *exemplar*, which is simply an image of a single instance of that object. Recently, counting models have become even more accessible with models for which the object to be counted is specified via a text prompt.

However, we observed that recent models such as CountSE (Liu et al., 2025) and CountGD (Amini-Naieni et al., 2024), that perform very well when objects are easily distinguishable and separated, struggle significantly in many real-world settings illustrated in Fig. 2.

We hypothesize that the cause lies principally in the training data. In addition to being severely limited in both scale and diversity, manual annotation at high count produces systematic errors, as shown in Fig. 3. These errors concentrate in exactly the cluttered, high-count images that are most valuable for training, hindering the accuracy of any model trained on them. Furthermore, existing synthetic datasets do not have the same scale limitation, but they greatly lack diversity due to their simplistic design. This results in an important domain gap between training samples and real-world images, limiting the generalization of the downstream model. Finally, constructing training data by prompting a diffusion model to generate a predefined number of objects is ineffective due to their inability to synthesize images with a controlled count (Kajić et al., 2024).

In parallel, 3D reconstruction has made significant advances in recent years, leading to increasingly realistic and widespread real-world captures of objects, materials, and lighting environments. These advancements make it possible to generate arbitrary numbers of synthetic scenes with the photorealism of real data, since the components of the scenes are from real-world captures themselves. This has been successfully exploited in recent datasets such as Kubric (Greff et al., 2022) for video tracking, HyperSim (Roberts et al., 2021) for 3D understanding, or OmniObject3D (Wu et al., 2023) for 3D generation. In this work, we aim to construct the first equivalent data generator for visual counting.

Therefore, our proposed solution to bridging the data gap is a complete data generation framework that takes as input real-world captures and outputs photorealistic samples that can be used to train a counting model. We design our generator to produce training data matching the current failure modes highlighted in Fig. 2, including repetitive backgrounds, visually-similar and self-similar objects, or size variations. The produced samples have pixel-perfect ground-truth, as well as unprecedented scale and diversity. We also expand the capabilities of our generator to provide several possible exemplars and text prompts for each object, making the downstream models more robust to the input prompt.

Finally, we release MixCount, the first dataset generated with this framework. MixCount includes 58,000 images featuring 1,522 different objects. In our experiments, augmenting the real training data of state-of-the-art counting models with MixCount led to a substantial reduction in model prediction error across modern benchmarks. At evaluation, MixCount also serves as a powerful diagnostic benchmark. Its exact, dense annotations of object visibility, size, exemplars, and text prompts let us

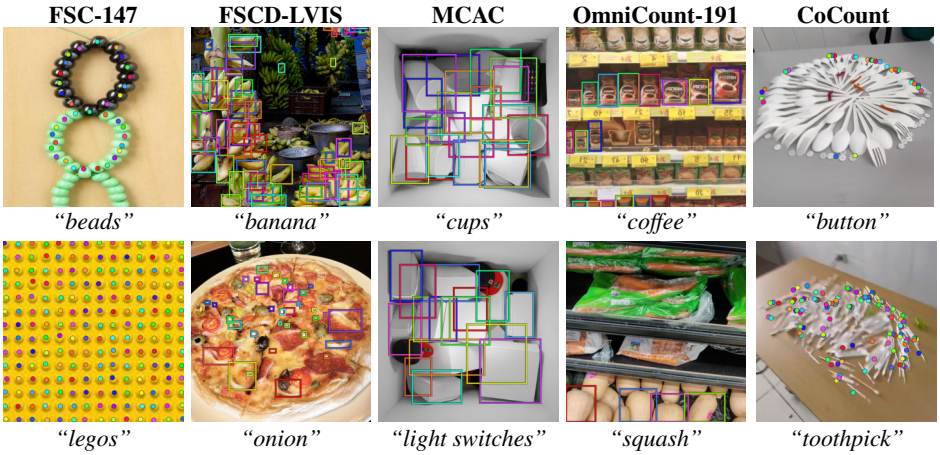


Figure 3: **Ground-truth annotations of existing counting datasets.** Previous datasets all exhibit significant annotation errors, motivating the need for a training and benchmarking dataset with guaranteed accurate annotations. FSC-147, FSCD-LVIS, OmniCount-191, and CoCount miss instances or have ambiguous descriptions, while MCAC includes strongly-occluded or even hidden objects.

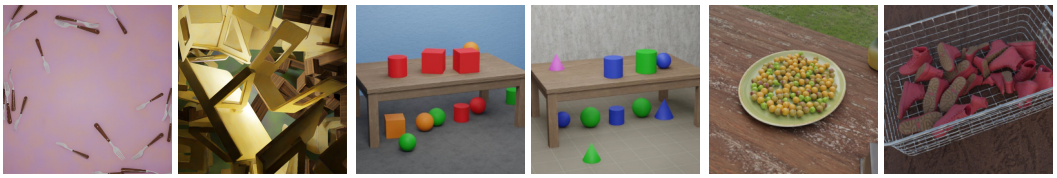
isolate how each factor affects counting error, which real-world benchmarks cannot do. In summary, our contributions are:

- The MixCount dataset, a large-scale, photorealistic, and diverse dataset with exact annotations to train and evaluate visual counting models,
- A data generation framework that synthesizes virtually infinite counting data for fine-grained object counting and can easily be tailored to specific applications,
- We demonstrate that models trained with MixCount achieve improved performance on real-world datasets, with 35-39% and 20.14% error reductions on the challenging PairTally (Nguyen et al., 2025) and FSC-147 (Ranjan et al., 2021) benchmarks, respectively.
- A diagnostic benchmark whose exact, dense annotations attribute the performance degradation of state-of-the-art counting models to specific factors such as object visibility, relative size, exemplar quality, and prompt detail.

2 Related Works

Visual counting models. Methods for automated object counting in images started as class-specific, capable of counting only one type of object. These include techniques for counting humans (Ranjan et al., 2018; Liu et al., 2019; Zhang et al., 2025) in crowds, predefined items in industrial inspection and retail (Jenkins et al., 2023; Dumery et al., 2026), animals in the wild (Arteta et al., 2016), and cells (Flaccavento et al., 2011; Guo et al., 2021). More recent *open-world* counting models can in theory count *any* specified object given both the input image and multi-modal object prompts. CounTX (Amini-Naieni et al., 2023), CLIP-Count (Jiang et al., 2023), and CountSE (Liu et al., 2025) accept text prompts that describe the object. Visual question answering (Antol et al., 2015; Zhang et al., 2018; Kim et al., 2025) and Referring Expression Counting (Dai et al., 2024; Wang et al., 2025; Zou et al., 2026) models rely on open-ended and contextual text prompts. These developments were enabled by the availability of text prompts for the FSC-147 and REC-8k datasets. CounTR (Liu et al., 2022), GeCo (Pelhan et al., 2024b), GeCo2 (Pelhan et al., 2026), DAVE (Pelhan et al., 2024a), and LOCA (Đukić et al., 2023) accept visual exemplars, in the form of bounding boxes over example instances. State-of-the-art models CountGD (Amini-Naieni et al., 2024) and CountGD++ (Amini-Naieni & Zisserman, 2026a) accept both visual exemplars and text as prompts. While this progress is promising, these techniques still struggle to distinguish between visually similar objects.

Counting datasets. The most commonly used datasets for training and evaluating counting models only have one category of object labeled per image, including cars (Mundhenk et al., 2016), humans (Sindagi et al., 2019; 2020), cells (Huang et al., 2020), fish (Kay et al., 2022), or surgery instruments (Bhyri et al., 2026). In open-world counting, the standard FSC-147 (Ranjan et al., 2021) dataset contains 6135 images with objects from 147 different categories, but only one type of object is



(a) MCAC (b) SITUATE (c) MixCount (Ours)

Figure 4: **Large-scale synthetic data comparison.** (a) MCAC (Hobley & Prisacariu, 2024) and (b) SITUATE (Peinl et al., 2026) samples lack the complexity of real-world images. In comparison, (c) MixCount samples are more diverse, realistic, and challenging.

labeled per image. Similarly, FSCD-LVIS (Nguyen et al., 2022), as well as many video (Amini-Naieni & Zisserman, 2026b) and stacked object counting datasets (Dumery et al., 2025) label only one category per sample. This has embedded a bias in SoTA counting models trained and evaluated on these datasets, which struggle to distinguish between visually similar instances.

Recent work introduces new but limited real-world datasets with multiple categories of objects labeled. Figure 3 shows some of the annotation errors and limitations of existing datasets. The PairTally (Nguyen et al., 2025) test set includes 681 images of paired objects that are mixed together with ground truth counts, visual exemplars, and text descriptions. The OmniCount-191 (Mondal et al., 2025) dataset consists of 30,230 real-world images. The REC-8k (Dai et al., 2024) dataset includes 8000 images with contextual text descriptions, indicating relative positions or attributes. The Pixmo-Count (Deitke et al., 2025) dataset contains images, text descriptions, and point annotations for 36,900 training images and 540 validation and test images. The Lookalikes test set (D'Alessandro et al., 2025) contains images, text descriptions, and ground truth counts. PairTally and the Lookalikes datasets lack training sets, so they are not suitable for training counting models. Finally, CoCount (Huang et al., 2026) is a training set in which only two categories are labeled per-image. CoCount is nonetheless restricted to the same categories as PairTally, and its frames are sampled from twenty videos per category pair, with dot labels propagated across frames by an off-the-shelf tracker, which limits scene diversity and introduces annotation errors. REC-8k and Pixmo-Count lack exemplars, so they cannot train exemplar-based counters. General-purpose detection datasets do not fill the gap either, with MS COCO (Lin et al., 2014) lacking annotations beyond 15 instances of a category per image.

To address the limitations of real-world data, recent work has turned to synthetic data. The GCC (Wang et al., 2019) dataset leveraged video game assets to produce synthetic crowd counting samples. The MCAC (Hobley & Prisacariu, 2024) dataset features high counts of mixed objects and ground truth bounding boxes. The synthetic SITUATE (Peinl et al., 2026) dataset contains images with only four object classes: cubes, spheres, cones and cylinders. Both MCAC and SITUATE lack photorealism. MCAC and StackCounting (Dumery et al., 2025) both lack text descriptions and count hidden objects, making them unsuitable for 2D multi-modal counting. We compare our MixCount dataset to these synthetic alternatives in Fig. 4 and in our experiments.

Overall, existing datasets lack a combination of (i) multiple visually similar object categories labeled per image, (ii) accurate multi-modal prompts and annotations, and (iii) scalable, realistic, and diverse training data, restricting progress in object counting. In contrast, MixCount provides 58,000 images spanning 1,522 object classes, each object carrying a pixel-perfect bounding box, segmentation mask, exemplars and several text descriptions.

3 The MixCount Dataset and Data Generator

3.1 Large-scale counting dataset with exact annotations

Diversity and scale. With this work, we release the initial version of the MixCount dataset, the first large-scale photorealistic synthetic dataset for training and benchmarking visual counting models. MixCount represents a significant leap in the diversity of richly annotated counting datasets, as can be seen in Fig. 4. Overall, MixCount includes 58,000 images containing a total of 4 million objects to count. There are 68 objects per-scene on average sampled from two to four different classes. This encompasses 1,522 different commonly found real-world objects spanning a large number of categories. We provide additional statistics on the dataset and objects in our supplementary.

Image	Exemplars			Descriptions		
				Short	Concise	Detailed
	 ext.	 s=1.0	 s=0.60	gold paneled volleyball	Shiny golden volleyball with dark seams	Shiny golden volleyball features segmented panels joined by dark seams, displaying a smooth and reflective surface.
	 ext.	 s=1.0	 s=0.52	yellow white volleyball	Yellow and white paneled volleyball with text	Volleyball composed of alternating yellow and white panels, with 'VOLLEYBALL soft touch' printed in black lettering on a yellow section.

Figure 5: **Dataset features.** MixCount includes different exemplars and tiered description granularity.

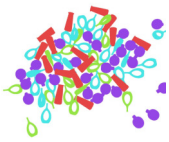
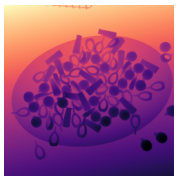
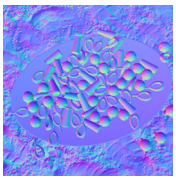
Image	Instance Segmentation	Class Segmentation	Bounding Boxes	Depth	Normal Map
					

Figure 6: **Dense ground-truth annotations.** Each sample is annotated with object localization labels, as well as depth and normal maps.

Object prompts. To allow the training and evaluation of recent multimodal models, each object in MixCount is associated with both text descriptions and visual exemplars. These *object prompts* can be used to specify which object is to be counted and we display some examples in Fig. 5. Regarding text descriptions, we observed that previous datasets often suffer from simplistic or ambiguous descriptions, such as *bird*, *lego*, *finger food* in FSC-147. As a consequence, the counting models trained on these datasets generalize poorly to real-world applications, in which several qualifiers are often necessary to designate a specific object. To remedy this, each object in MixCount is annotated at three different levels: short, concise and detailed descriptions. These serve as unique identifiers and we manually ensure that no two objects have an identical text prompt.

We also introduce two different kinds of visual exemplars. An *internal* exemplar, represented as a bounding box in the image, and an *external* exemplar, which is an image of the same object but in a different setting. Internal exemplars are assigned a score s that relates to the underlying object's visibility and size in the image. MixCount also includes external exemplars of each object in their canonical pose in front of a white background, as well as crops from different scenes of our dataset. In the supplementary, we measure the influence of the different exemplars and text descriptions.

Annotations. Every object is annotated with a pixel-perfect bounding box and segmentation mask. These annotations are saved in COCO format and can directly be integrated in the majority of existing pipelines to train visual counting models. By design, all objects in the ground-truth annotations are guaranteed to be visible in that image, which was an issue in previous datasets such as MCAC where objects were often completely hidden, leading to inconsistent supervision in training. Finally, we also save segmentations of the objects, ground-truth depth and normal maps, as illustrated in Fig. 6.

Splits. We divide the MixCount dataset into train-val-test split, including 56,000, 1000 and 1000 images, respectively. To guarantee that no object found in the test set can be seen in the train set, we split the objects to be counted directly, such that each object can only be found in a single split. To ensure compatibility with models trained and evaluated on FSC-147, we match each object in our dataset with a category in FSC-147 when applicable and use the same train-test split. For example, since *ball* is a train class in FSC-147, all types of balls such as *Light orange football* or *Camo patterned basketball* are part of the train split. Finally, the objects which do not broadly match any of FSC-147's limited classes are then assigned a random split.

3.2 Data Generator

Our central motivation is to provide counting data with exact annotations that are as photorealistic, challenging and diverse as possible. Thus, we design the MixCount data generator to address the

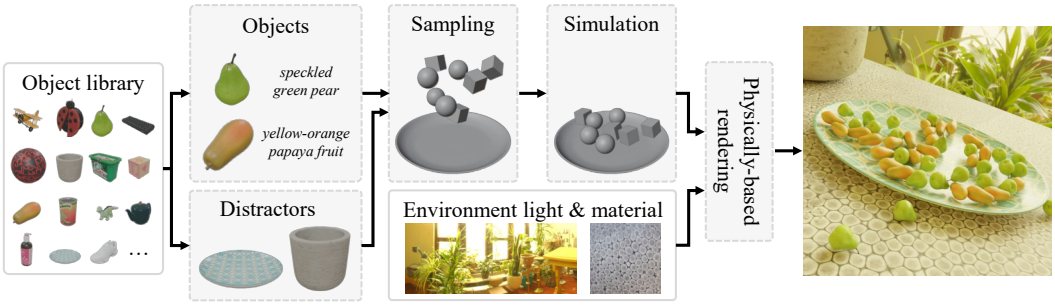


Figure 7: **Data generator.** Our generator samples objects, distractors, environment and camera placement to procedurally generate photorealistic training samples. All assets are issued from high-quality captures of real-world objects.

failure modes of current counting models in a fully automatic fashion and at scale. We opt for a synthetic data generator depicted in Figure 7 that incorporates captures of real-world objects, environments, and materials, and physically simulates them to compose realistic scenes. In combination with physically-based rendering, this provides our synthesized samples the photorealism and complexity of real-world images while maintaining the annotation quality and scale of synthetic data. Due to space constraints, further details on our data generator, visualization of external exemplars, and dataset statistics can be found in our supplementary.

Objects. For each scene, we start by sampling up to four different object classes. Objects are randomly picked from the Digital Twin Catalog (DTC) (Dong et al., 2025), a dataset of high-quality captures of real-world objects. We manually inspect the dataset to filter out objects that are too similar to be distinguished by text or visual exemplar. After the first object class is created, the subsequent classes are generated either by modifying its size, by sampling another object from the same DTC category, or by sampling from a different category with a reduced probability. This favors meaningful combinations while preserving the possibility to generate improbable mixes adding to data diversity.

Simulation. We sample 3D positions and spawn an object from a random class at each position. A slight scale modifier is applied to each object, ensuring all objects of the same class are different while remaining visually similar. We then employ the physical simulation of Blender (Blender Online Community, 2018) to let the objects naturally form a realistic scene, with objects touching and partially occluding each other.

Occlusions. Previous dataset generation pipelines provide no guarantee that all objects to be counted are indeed visible. MixCount solves this by explicitly measuring the visibility of each object and removing occluded objects. Specifically, for each scene a segmentation image is rendered from which masks for all N objects in the scene can be extracted. However, these masks only cover parts of the objects since other objects or elements of the environment may occlude them partially or completely. We will thus refer to these masks as the *occluded* masks $(M_{o,i})_{1\leq i\leq N}$. To remedy this, we also render *unoccluded* masks $(M_{u,i})_{1\leq i\leq N}$, in which all occlusions are ignored, by doing a fast render pass per-object with all other objects disabled. This allows our data generator to measure, for an object i, the visibility ratio $v_i = \frac{|M_{o,i}|}{|M_{u,i}|}$. Given an input visibility threshold v_t, the occluded objects with $v_i < v_t$ are thus removed from the scene prior to rendering. Since doing this only makes the remaining objects more visible, this guarantees that all remaining objects are visible by at least v_t.

Environment. Models trained on existing datasets fail to generalize to the rich and challenging environments found in real-world images, partly due to their simplistic backgrounds. MixCount addresses this issue with procedural generation, real object captures and physically-based rendering. All assets used to produce MixCount are based on real-world captures and are parts of publicly released datasets, ensuring the complete reproducibility of our dataset.

First, we employ real-world captures of environment maps to provide a complex and diverse lighting environment, both in the directionality of light sources and their intensity. For each scene, a random environment map is sampled from the 2233 available in the LavalIndoor (Gardner et al., 2017) dataset or the 963 in Polyhaven (Zaal et al., 2021). We also use 568 physically-based rendering materials (PBR) from Polyhaven and 150 from VasTexture (Eppel, 2024) to texture the floor upon which the objects lie. We purposefully favor repetitive textures that are challenging for counting models and are absent in previous datasets.

Finally, we introduce *distractors* to the scene, which are objects from the DTC dataset just like the main objects to count. However, these objects do not repeat and only serve to add complexity to the scene and cast shadows on the main objects. Distractors are guaranteed to be different objects from the main objects to count, and their bounding boxes are also reported in the final results.

Rendering. In each simulated scene, we define the camera origin through its azimuthal and polar angles (ϕ, θ) as well as its radius r, which are sampled from an acceptable range. The camera rotation is dynamically set to look at the center of the objects to count, then slightly jittered to introduce diversity and mimic an imperfect capture. Finally, we generate an image through ray-tracing, which computationally simulates reflections in the scene and produces photorealistic renders.

Exemplars and text descriptions. Internal exemplars are extracted from the segmentation maps. To provide external exemplars, we also render each DTC object under a standard lighting and with a white background. We task a vision-language model with captioning these images into short, concise, and detailed descriptions for each object. Following this, we remove all objects with identical descriptions to ensure each object is uniquely distinguishable. Text descriptions are manually audited.

Customization. We designed the MixCount data generator to be highly versatile and easily tailored for a specific application. For instance, in the case of industrial inspection where specific boxes and items are present in a warehouse, it is now fairly straightforward with image-to-3D or text-to-3D methods (Xiang et al., 2025b;a) to generate 3D assets of the specific items of interest. These assets can then be plugged into the MixCount data generator to synthesize a large-scale dataset tailored for this application, with a custom number of objects generated and multi-view camera placement.

4 Experiments

We demonstrate that MixCount can be used to improve the counting accuracy of state-of-the-art counting models on real-world data in Section 4.2, and evaluate these models against the challenging MixCount benchmark in Section 4.4. The realism, performance improvement at different scales, and the domain gap of our dataset are evaluated in Section 4.3. Additional results, including sensitivity studies to the choice of exemplar or text prompt and the influence of occlusion on counting accuracy are provided in the supplementary, as well as detailed training and evaluation protocols.

4.1 Experimental setup

Datasets. In addition to MixCount, four real-world counting datasets are used. FSC-147 (Ranjan et al., 2021) is the standard open-world counting dataset covering 147 classes and 6135 images with 7-3731 objects per image. PairTally (Nguyen et al., 2025) is a test set for fine-grained counting with 681 dense images. The SHA (Zhang et al., 2016) crowd-counting dataset is added to measure transfer to a completely different domain. We also report results on FSCD-LVIS (Nguyen et al., 2022).

Metrics. Following prior work, we report the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). On MixCount, we additionally report the Normalized Absolute Error (NAE) and the Accuracy 10% (Acc10%) defined in the supplementary.

Implementation. We train GeCo2 (Pelhan et al., 2026) and CountGD++ (Amini-Naieni & Zisserman, 2026a) on MixCount in various settings, including augmenting existing datasets with 1k MixCount images, and training on MixCount images alone. For evaluating models on MixCount, we use the official published pretrained checkpoints. Further details are in the supplementary.

4.2 Results - MixCount value for training

In Fig. 2, we show that training on MixCount remedies common failure modes of state-of-the-art counting models. In (a), MixCount enables CountGD++ to distinguish between differences in object size. In (b), the improved model identifies each pair of sunglasses as a single unit, instead of counting each self-similar lens, a well-known issue with counting models (Amini-Naieni et al., 2023). In (c), MixCount allows CountGD++ to ignore repetitive backgrounds and focus on the specified object.

Improving in-domain performance. In Table 1, training CountGD++ on a mix of real FSC-147 images and only 1k MixCount images achieves state-of-the-art accuracy on FSC-147, reducing the MAE by 20%. Using 1k MixCount images ensures a balance between real and synthetic data.

Model	Training data	FSC-147 Test	
		MAE ↓	RMSE ↓
DAVE	FSC-147 [real]	8.66	32.36
GeCo	FSC-147 [real]	7.91	54.28
GeCo2	FSC-147 [real]	7.64	39.39
CountGD	FSC-147 [real]	5.74	24.09
CountGD++	FSC-147 [real]	5.86	18.79
CountGD++	FSC-147 [real] + 1k MCAC [synthetic]	8.32	89.61
CountGD++	FSC-147 [real] + 1k OmniCount [real]	11.77	124.76
CountGD++	**FSC-147 [real] + 1k MixCount [synthetic] (ours)**	**4.68**	**15.66**

Table 1: **Improving in-domain performance on FSC-147.** Adding 1k MixCount images to FSC-147 improves counting accuracy on FSC-147, while using alternative mixed-category data does not. Exemplar-based models, DAVE, GeCo, and GeCo2 are given 3 positive visual exemplars to specify the target object. As multi-modal models, CountGD and CountGD++ are also given text.

In Table 1, we also show replacing the MixCount images with alternative multi-category data harms in-domain performance. 1k images are randomly sampled from MCAC (Hobley & Prisacariu, 2024) and OmniCount (Fruits) (Mondal et al., 2025) for training. Only the Fruits subset is sampled as it has been deemed accurate by prior works (Amini-Naieni & Zisserman, 2026a). MCAC raises the MAE from 5.86 to 8.32, and OmniCount-191 raises it to 11.77, while MixCount lowers it to 4.68.

Improving out-of-domain performance. In Table 2, we show that training GeCo2 and CountGD++ on MixCount leads to improved generalization to new datasets. Training GeCo2 on 56k MixCount images results in a 30% lower MAE on PairTally than training on the real FSC-147 training set. Training on a mix of images from both results in a 39% reduction of the MAE, achieving the best performance for GeCo2. Training CountGD++ on a mix of FSC-147 and MixCount results in a 35% reduction in the MAE on PairTally, and a small improvement on SHA. On the other hand, adding OmniCount images increases the MAE on both datasets. Adding MCAC images improves the error on PairTally, but slightly increases it on SHA. MixCount reduces error more than MCAC on PairTally.

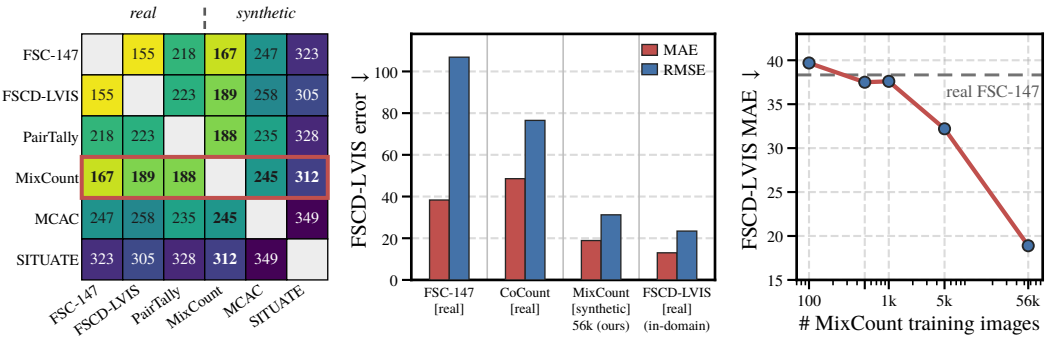
Model	Training data	PairTally Test		SHA Test	
		MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
DAVE	FSC-147 [real]	47.37	84.31	-	-
GeCo	FSC-147 [real]	50.24	73.41	-	-
GeCo2	FSC-147 [real]	52.36	76.46	-	-
GeCo2	56k MixCount [synthetic] (ours)	35.85	64.89	-	-
GeCo2	**FSC-147 [real] + 1k MixCount [synthetic] (ours)**	**31.88**	**52.84**	-	-
CountGD++	FSC-147 [real]	35.27	60.85	116.00	234.00
CountGD++	FSC-147 [real] + 1k MCAC [synthetic]	32.48	50.00	116.39	237.95
CountGD++	FSC-147 [real] + 1k OmniCount [real]	43.66	73.66	203.60	352.32
CountGD++	**FSC-147 [real] + 1k MixCount [synthetic] (ours)**	**22.94**	**45.08**	**113.60**	**225.31**

Table 2: **Improving out-of-domain performance on PairTally and SHA.** Adding 1k MixCount images to FSC-147 improves counting accuracy on PairTally and SHA. Alternative mixed-category data do not lead to the same level of improvement. DAVE and GeCo variants are given 3 positive visual exemplars. CountGD++ variants are given both 3 positive and negative visual exemplars and text for PairTally. SHA only allows for text inputs. - indicates methods that do not allow for text input.

4.3 Results - Dataset analysis

Proximity to real data. Figure 8a reports the Fréchet-Inception Distance between pairs of counting datasets. MixCount is the synthetic dataset closest to all three real ones, and is closer to FSC-147 and PairTally than those two are to each other. Error on the MixCount test set also correlates with error on real PairTally across the models we evaluated on both with a correlation of $r \approx 0.86$, suggesting that our benchmark is a useful proxy for real-world counting difficulty.

MixCount-only training. In Fig. 8b, we evaluate the domain gap between MixCount and the real-world FSCD-LVIS dataset and compare it to the gap with FSC-147 and CoCount. GeCo2 trained on 56k MixCount images achieves an MAE that over 50% lower than the MAE from training on the real FSC-147 and CoCount datasets. As expected, the error on FSCD-LVIS from training only on MixCount is still not as low as the error achieved by training on FSCD-LVIS itself.



(a) FID Feature proximity (b) Synthetic-only training (c) Scaling curve

Figure 8: **Dataset analysis.** We quantify the realism of MixCount images (a), the out-of-domain accuracy reached without any real training data (b), and the impact of data scale (c).

Scaling with data volume. In Fig. 8c, we train a GeCo2 model on MixCount images alone and gradually increase the number of training samples. Error on FSCD-LVIS decreases steadily, and roughly 1k synthetic images already match a model trained on the entire real FSC-147 training set.

4.4 Results - MixCount value for benchmarking

Evaluating in the mixed-object setting. In Table 3, we show existing state-of-the-art counting models CountSE (Liu et al., 2025), CountGD, and CountGD++ trained on FSC-147 perform poorly in the mixed-object setting. The highest Acc10% of 0.33 is achieved by CountGD++ given both positive and negative prompts. As expected, training on MixCount improves the Acc10% to 0.69 and the NAE to 0.14. Additional details on our benchmarking protocol and many qualitative examples of applying CountGD++ (+ MixCount) to the MixCount test set are in the supplementary.

Model	Prompt				MAE ↓	RMSE ↓	NAE ↓	Acc10% ↑
	t^+	B^+_{int}	t^-	B^-_{int}				
CountSE	✓	✗	✗	✗	32.49	41.49	7.23	0.06
CountGD	✓	✗	✗	✗	33.00	42.25	5.64	0.05
CountGD++	✓	✗	✗	✗	26.77	36.20	2.73	0.09
CountGD	✓	✓	✗	✗	41.83	54.42	11.09	0.01
CountGD++	✓	✓	✗	✗	34.55	46.57	4.60	0.06
CountGD++	✓	✓	✓	✓	8.12	22.83	1.21	0.33
CountGD++ (+ MixCount)	✓	✗	✗	✗	14.63	22.14	1.02	0.09
CountGD++ (+ MixCount)	✓	✓	✗	✗	12.16	20.08	1.02	0.23
CountGD++ (+ MixCount)	✓	✓	✓	✓	**2.11**	**4.73**	**0.14**	**0.69**

Table 3: **Benchmarking existing models on our MixCount test set.** CountGD++ (+MixCount) is the model trained on FSC-147 + 1k MixCount. t^+, B^+_{int}, t^-, B^-_{int} stand for positive text, 3 positive exemplars, negative text, and 3 negative exemplars respectively.

Additional evaluations. Due to its rich and accurate annotations and breadth of prompts, MixCount enables extensive and reliable evaluation of counting models. In the supplementary, we evaluate CountGD++'s sensitivity to **internal vs. external exemplars**, **exemplar quality**, **target object visibility**, and the **descriptiveness of the text prompt**.

5 Conclusion

We have released the MixCount dataset, the first synthetic, large-scale, photorealistic and diverse training and evaluation dataset for open-vocabulary counting constructed from captures of real-world objects. Adding MixCount images into real-world training sets can significantly improve counting performance. MixCount's new features, including external exemplars and textual descriptions at varying levels of detail, make future work on prompt sensitivity and versatility possible. Because its annotations are exact and dense, the degradation it exposes can be attributed to specific factors such as object visibility, relative size, exemplar quality and prompt detail, which no manually annotated counting dataset permits. Finally, our versatile and fully automated data generator synthesizes perfectly labeled and diverse counting data at scale.

AI USE STATEMENT

We used generative AI tools as a component of our data generation pipeline. The three levels of object description released with MixCount are produced by a vision-language model prompted with renderings of each 3D asset. The resulting annotations are part of the released artifact. Descriptions were manually verified by the authors against a rendering of the corresponding object, as described in Section C. The small number that were inaccurate led to the corresponding asset being discarded. We additionally used generative AI tools for editing assistance, for data processing and filtering, for producing visualisations of our results, for assistance in running experiments, and for implementing standard methods. All AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was read and revised by the authors. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work releases a dataset, and its principal ethical considerations concern the content of that dataset and the uses it enables. MixCount contains no humans, no personal data and no identifiable content. Every scene is rendered from captures of inanimate objects. Restricting the dataset to non-human categories limits the most direct misuse of an accurate counting model, namely surveillance of people.

REPRODUCIBILITY STATEMENT

The MixCount dataset, the generator source code, and all text annotations used in our experiments will be publicly released. The appendix documents the complete set of simulation parameters, including scene construction, object sampling, physics, camera sampling, rendering settings and visibility filtering, which together determine the data distribution. Training and evaluation hyperparameters are also listed in our supplementary.

REFERENCES

Niki Amini-Naieni and Andrew Zisserman. CountGD++: Generalized prompting for open-world counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Niki Amini-Naieni and Andrew Zisserman. Open-world object counting in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2300–2308, 2026b.

Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specified object counting. In *The 36th British Machine Vision Conference (BMVC)*, 2023.

Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. CountGD: Multi-modal open-world counting. *Advances in Neural Information Processing Systems*, 37:48810–48837, 2024.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European conference on computer vision*, pp. 483–498. Springer, 2016.

Rishikesh Bhyri, Brian R Quaranto, Junsong Yuan, Peter CW Kim, and Nan Xi. Chain-of-look spatial reasoning for dense surgical instrument counting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8521–8530, 2026.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Siyang Dai, Jun Liu, and Ngai-Man Cheung. Referring expression counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16985–16995, 2024.

Adrian D'Alessandro, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Figo: Fine-grained object counting without annotations. *arXiv preprint arXiv:2504.11705*, 2025.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems*, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, et al. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 753–763, 2025.

Corentin Dumery, Noa Etté, Aoxiang Fan, Ren Li, Jingyi Xu, Hieu Le, and Pascal Fua. Counting stacked objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19774–19783, 2025.

Corentin Dumery, Noa Etté, Aoxiang Fan, Ren Li, Jingyi Xu, Hieu Le, and Pascal Fua. Automated counting of stacked objects in industrial inspection. *arXiv preprint arXiv:2603.15470*, 2026.

Sagi Eppel. VasTextures: Vast repository of textures and pbr materials extracted from real-world images using unsupervised methods. *arXiv preprint arXiv:2406.17146*, 2024.

G. Flaccavento, Victor Lempitsky, Iestyn Pope, P. R. Barber, Andrew Zisserman, J. Alison Noble, and B. Vojnovic. Learning to count cells: applications to lens-free imaging of large fields. In *Microscopic Image Analysis with Applications in Biology*, 2011.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36(6):1–14, 2017.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3749–3761, 2022.

Yue Guo, Oleh Krupa, Jason Stein, Guorong Wu, and Ashok Krishnamurthy. Sau-net: A unified network for cell counting in 2d and 3d microscopy images. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(4):1920–1932, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Michael Hobley and Victor Prisacariu. Abc easy as 123: A blind counter for exemplar-free multi-class class-agnostic counting. In *European Conference on Computer Vision*, pp. 304–319. Springer, 2024.

Yifeng Huang, Gia Khanh Nguyen, and Minh Hoai. CountEx: Fine-grained counting via exemplars and exclusion. *arXiv preprint arXiv:2602.19432*, 2026.

Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Ruizhe Geng, Hongliang He, Shan Du, Xia Liu, Yonghong Tian, Yongsheng Liang, et al. Bcdata: A large-scale dataset and benchmark for cell detection and counting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 289–298. Springer, 2020.

Porter Jenkins, Kyle Armstrong, Stephen Nelson, Siddhesh Gotad, J Stockton Jenkins, Wade Wilkey, and Tanner Watts. Countnet3d: A 3d computer vision approach to infer counts of occluded objects. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3008–3017, 2023.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. CLIP-Count: Towards text-guided zero-shot object counting. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4535–4545, 2023.

Ivana Kajić, Olivia Wiles, Isabela Albuquerque, Matthias Bauer, Su Wang, Jordi Pont-Tuset, and Aida Nematzadeh. Evaluating numerical reasoning in text-to-image models. *Advances in neural information processing systems*, 37, 2024.

Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. In *European Conference on Computer Vision*, pp. 290–311. Springer, 2022.

Aniket A Khule, Manoj S Nagmode, and Rajkumar D Komati. Automated object counting for visual inspection applications. In *2015 International Conference on Information Processing (ICIP)*, pp. 801–806. IEEE, 2015.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Computing Surveys*, 57(10):1–35, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Chang Liu, Yujie Zhong, Andrew Zisserman, and Weidi Xie. CounTR: Transformer-based generalised visual counting. *arXiv preprint arXiv:2208.13721*, 2022.

Shuai Liu, Peng Zhang, Shiwei Zhang, and Wei Ke. CountSE: Soft exemplar open-set object counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21536–21546, 2025.

Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5099–5108, 2019.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Anindya Mondal, Sauradip Nag, Xiatian Zhu, and Anjan Dutta. OmniCount: Multi-label object counting with semantic-geometric priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19537–19545, 2025.

T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European conference on computer vision*, pp. 785–800. Springer, 2016.

Gia Khanh Nguyen, Yifeng Huang, and Minh Hoai. Can current ai models count what we mean, not what they see? a benchmark and systematic evaluation. In *2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8. IEEE, 2025.

Thanh Nguyen, Chau Pham, Khoi Nguyen, and Minh Hoai. Few-shot object counting and detection. In *European Conference on Computer Vision*, pp. 348–365. Springer, 2022.

René Peinl, Vincent Tischler, Patrick Schröder, and Christian Groth. SITUATE–synthetic object counting dataset for VLM training. In *21st International Conference on Computer Vision Theory and Applications*, 2026.

Jer Pelhan, Alan Lukežic, Vitjan Zavrtanik, and Matej Kristan. DAVE - a detect-and-verify paradigm for low-shot counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23293–23302, June 2024a.

Jer Pelhan, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation. *Advances in Neural Information Processing Systems*, 37:66260–66282, 2024b.

Jer Pelhan, Alan Lukežič, and Matej Kristan. Generalized-scale object counting with gradual query aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8314–8321, 2026.

Mary C Phelan and Gretchen Lawler. Cell counting. *Current protocols in cytometry*, (1):A–3A, 1997.

Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 270–285, 2018.

Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3394–3403, 2021.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1221–1231, 2019.

Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2594–2609, 2020.

Nikola Đukić, Alan Lukežič, Vitjan Zavrtanik, and Matej Kristan. A low-shot object counting network with iterative prototype adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18872–18881, 2023.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8198–8207, 2019.

Zhicheng Wang, Zhiyu Pan, Zhan Peng, Jian Cheng, Liwen Xiao, Wei Jiang, and Zhiguo Cao. Exploring contextual attribute density in referring expression counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19587–19596, 2025.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. OmniObject3D: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 803–814, 2023.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025b.

Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292, 2018.

Greg Zaal, Rob Tuytel, Rico Cilliers, James Ray Cock, Andreas Mischok, Sergej Majboroda, Dimitrios Savva, and Jurita Burger. Polyhaven: a curated public asset library for visual effects artists and game designers, 2021.

Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 833–841, 2015.

Qi Zhang, Yunfei Gong, Zhidan Xie, Zhizi Wang, Antoni B Chan, and Hui Huang. Semi-supervised multi-view crowd counting by ranking multi-view fusion models. *arXiv preprint arXiv:2512.16243*, 2025.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi 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.

Yuda Zou, Zijian Zhang, and Yongchao Xu. Decoupling what to count and where to see for referring expression counting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14113–14121, 2026.