CVPR #360

Towards Universal Dataset Distillation via Task-Driven Diffusion

Supplementary Material

The supplementary material is organized as follows: Section 6 presents more experiments and analysis to this work; Section 7 provides more implementation details; Section 8 provides the GPT-4o generation prompts in Task-Aligned Contextual Prompting; and finally, Section 9 presents more synthetic images visualization.

6. More Experiments and Analysis

6.1. Experiment on CIFAR-10 and CIFAR-100

As shown in the Tab 7, previous optimization-based methods have predominantly been evaluated on small datasets such as CIFAR-10 and CIFAR-100. These datasets have a resolution of just 32×32, which gives optimization methods, particularly bilevel optimization approaches like DSA and TESLA, a distinct advantage. Diffusion-based methods, on the other hand, often struggle on such small datasets because the quality of image generation at low resolutions tends to be unstable. For our UniDD, the base SD V1.5 model is designed to generate images at a resolution of 512×512, and its direct generation of ultra-low resolution images is poor, often resulting in artifacts. To address this, we employ downsampling to align with the resolution of the CIFAR datasets, though this naturally leads to some performance degradation. Despite this, UniDD still performs well at IPC-50, achieving the second-best performance and outperforming the diffusion-based D^{4}M. However, low-resolution datasets are far from real-world applications, which is why we proceed to test on the more practical ImageNet dataset to further demonstrate the superiority of UniDD in this paper.

6.2. Increasing IPC Analysis

As shown in the Fig. 5, we compare the performance curves of different methods as IPC increases. Following the MiniMax Diffusion setting [14], using ResNet-18 as the model achieves higher performance compared to the results obtained with Conv-5 in the main text. The full dataset performance with ResNet-18 is 89.3, which serves as the theoretical upper bound. Our UniDD consistently maintains a significant lead, even at higher IPC values. Additionally, we observe that when IPC exceeds 50, optimization methods like DM and IDC perform worse than the baseline. In contrast, diffusion-based distillation methods continue to outperform random selection, highlighting the higher potential performance ceiling of diffusion-based approaches.

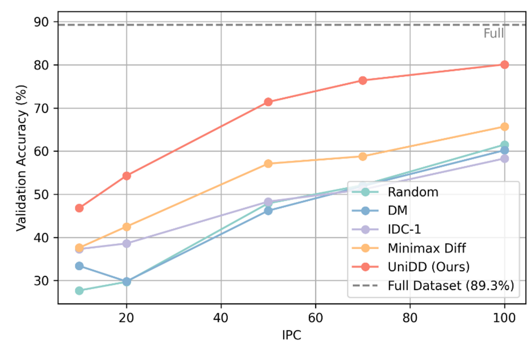

Figure 5. Validation Accuracy on ImageWoof: This shows the comparison with Random and other DD methods as the IPC increases. Test model is ResNet-18 and full dataset result is 89.3.

6.3. More Cross-Architecture Testing

In this section, we extend the cross-architecture experiments in classification to include detection and segmentation tasks, further validating the generalization ability of our synthetic datasets across different architectures. As shown in Table 2, we select Faster R-CNN and RetinaNet as the detectors, and LR-ASPP and FCN as the segmenters. These models are used interchangeably as the $\mathcal{T}_{\mathcal{SP}}$ model and the test model. Unlike traditional bilevel optimization methods, which often experience significant performance degradation when applied across architectures, our method demonstrates strong generalization. This is attributed to the diffusion-based generation process, which benefits stronger detectors and segmenters, showcasing the robustness and adaptability of our approach.

7. More Implementation Details

In the main experiments, we use the following models:
ResNet-18: A lightweight residual network with 18 layers, featuring residual connections for efficient feature learning.
Faster R-CNN: A two-stage detection framework with a ResNet-50 backbone and FPN for multi-scale feature representation.
LR-ASPP: A lightweight segmentation model using a MobileNetv3 backbone and an ASPP module for multi-scale context capture.

For the training process of the $\mathcal{T}_{\mathcal{SP}}$ models, we use the official pretrained weights from PyTorch: ResNet pre-

Dataset	IPC	FRePO	DSA	CAFE	IDM	TESLA	D^4M	UniDD (Ours)	Full
CIFAR-10	10	<u>65.5±0.6</u>	52.1±0.5	50.9±0.5	58.6±0.1	**66.4±0.8**	56.2±0.4	57.2±0.6	84.8±0.1
	50	71.7±0.2	60.6±0.5	62.3±0.4	67.5±0.1	72.6±0.7	<u>72.8±0.5</u>	**73.1±0.8**	
CIFAR-100	10	42.5±0.2	32.3±0.3	31.5±0.2	<u>45.1±0.1</u>	41.7±0.3	45.0±0.1	**45.3±0.4**	56.2±0.3
	50	44.3±0.2	42.8±0.4	42.9±0.2	**50.0±0.2**	47.9±0.3	48.8±0.3	<u>49.2±0.5</u>	

Table 7. Performance comparison with state-of-the-art methods on small datasets: CIFAR-10 and CIFAR-100. IPC refers to Images per Class. All methods use a 5-layer ConvNet as the test model. We train the network from scratch 5 times on the distilled dataset and evaluate them on the original test dataset to get the $\bar{x} \pm std$. The best results are marked in bold, while the second-best results are underlined.

Task	Test Model	$\mathcal{T}_{\mathcal{SP}}$ Model	
	Ratio: 1%	Faster R-CNN	RetinaNet
Det. / mAP (%)	Faster R-CNN	16.8±0.5	15.9±0.5
	RetinaNet	16.3±0.6	15.7±0.4
	Ratio: 3.5%	LR-ASPP	FCN
Seg. / mIoU (%)	LR-ASPP	10.3±0.5	10.9±0.4
	FCN	10.6±0.4	11.2±0.6

Task-Aligned Contextual Prompting

-Task: Classification
-Outputs: {Contextual prompts}

```
"An image of green apple."
"An image of pineapple."
"An image of banana."
"An image of strawberry."
"An image of orange."
```

Figure 6. Prompts generation for classification.

Task-Aligned Contextual Prompting

-Task: Object Detection & Segmentation
-Outputs: {Contextual prompts}

```
"A tray of hot dogs with ketchup
and mustard, placed on a table
next to a bowl of soda."
"An elephant is standing in a
fenced area with bushes and
trees."
"A man wearing a grey sweater and
sunglasses is sitting on a green
bench in a park."
"Two birds are sitting on a bird
feeder."
"A man and a boy are playing with
a light saber toy in a living
room."
"A double-decker bus is parked at
a bus stop."
```

Figure 7. Prompts generation for Object Detection and Segmentation.

trained on ImageNet-1k, Faster R-CNN on COCO, and LR-ASPP on PASCAL VOC.

For the training and testing phases of the synthetic dataset: **Classification tasks**: We align with the training and testing pipeline used in RDED [34]. **Detection tasks**: We follow the official parameter settings from MMDetection [4]. **Segmentation tasks**: We adopt the official parameter settings from MMSegmentation [6]. Additionally, since the synthetic dataset has a lower compression rate, standard training epochs may not be sufficient for convergence. To address this, we extend the number of epochs during training, referencing the epoch configurations used in MiniMax [14] for synthetic data training.

8. Task-Aligned Contextual Prompting

Classification. Classification task only focuses on identifying the primary object or category within the image. **Object Detection.** Object detection involves locating and identifying multiple objects in the image. **Semantic Segmentation.** Semantic segmentation provides pixel-level labels for distinct regions. The contextual prompts are the sentences chosen from a collection of 5 GPT-4o generated sentences.

9. More Visualization

In this section, we provide more visualizations of the synthetic images for classification, object detection and segmentation.

CVPR
#360

Figure 8. More visualizations selected from the distilled ImageFruit. The class names are marked at the left of each row.

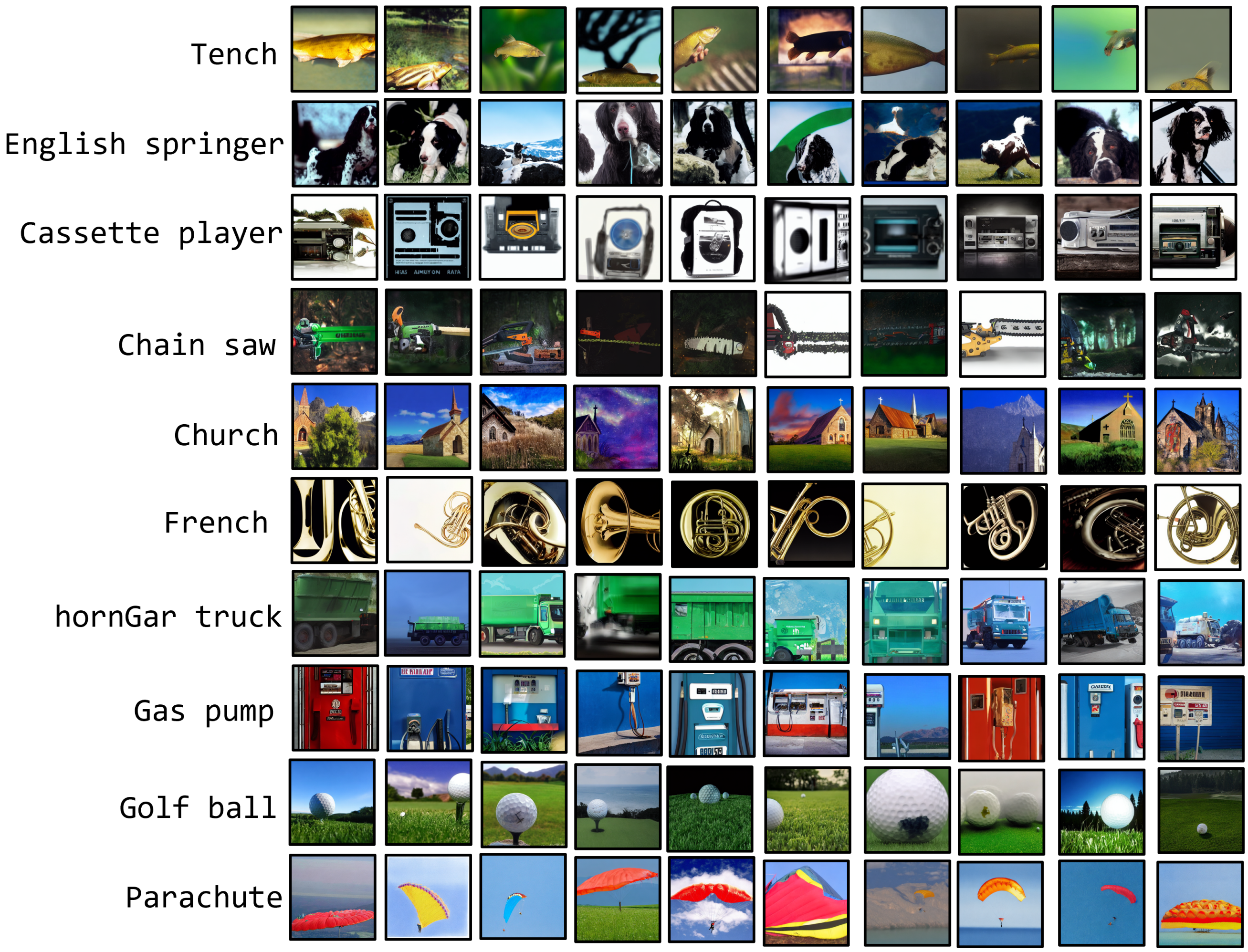

Figure 9. More visualizations selected from the distilled ImageNette. The class names are marked at the left of each row.

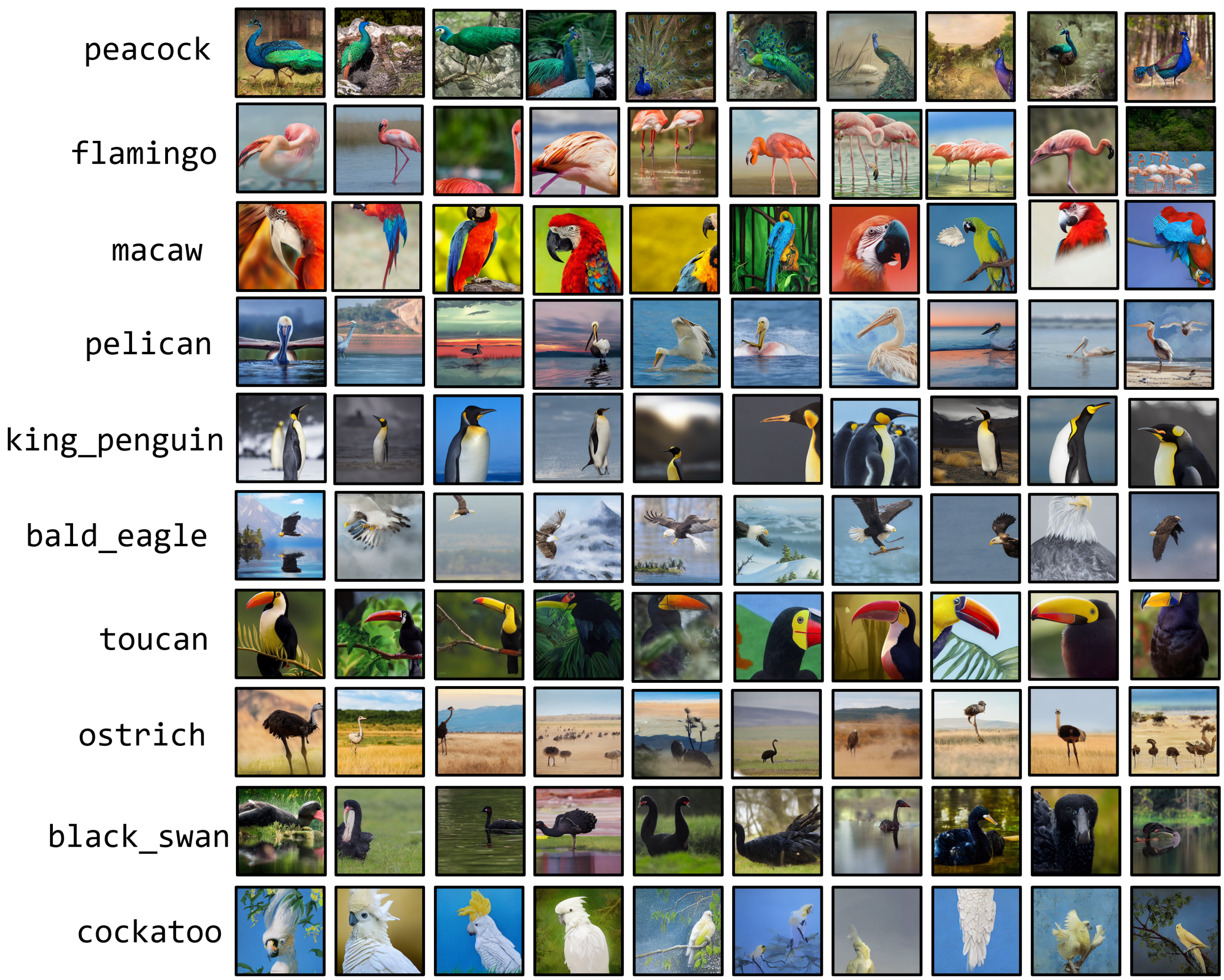

Figure 10. More visualizations selected from the distilled ImageSquawk. The class names are marked at the left of each row.

CVPR
#360

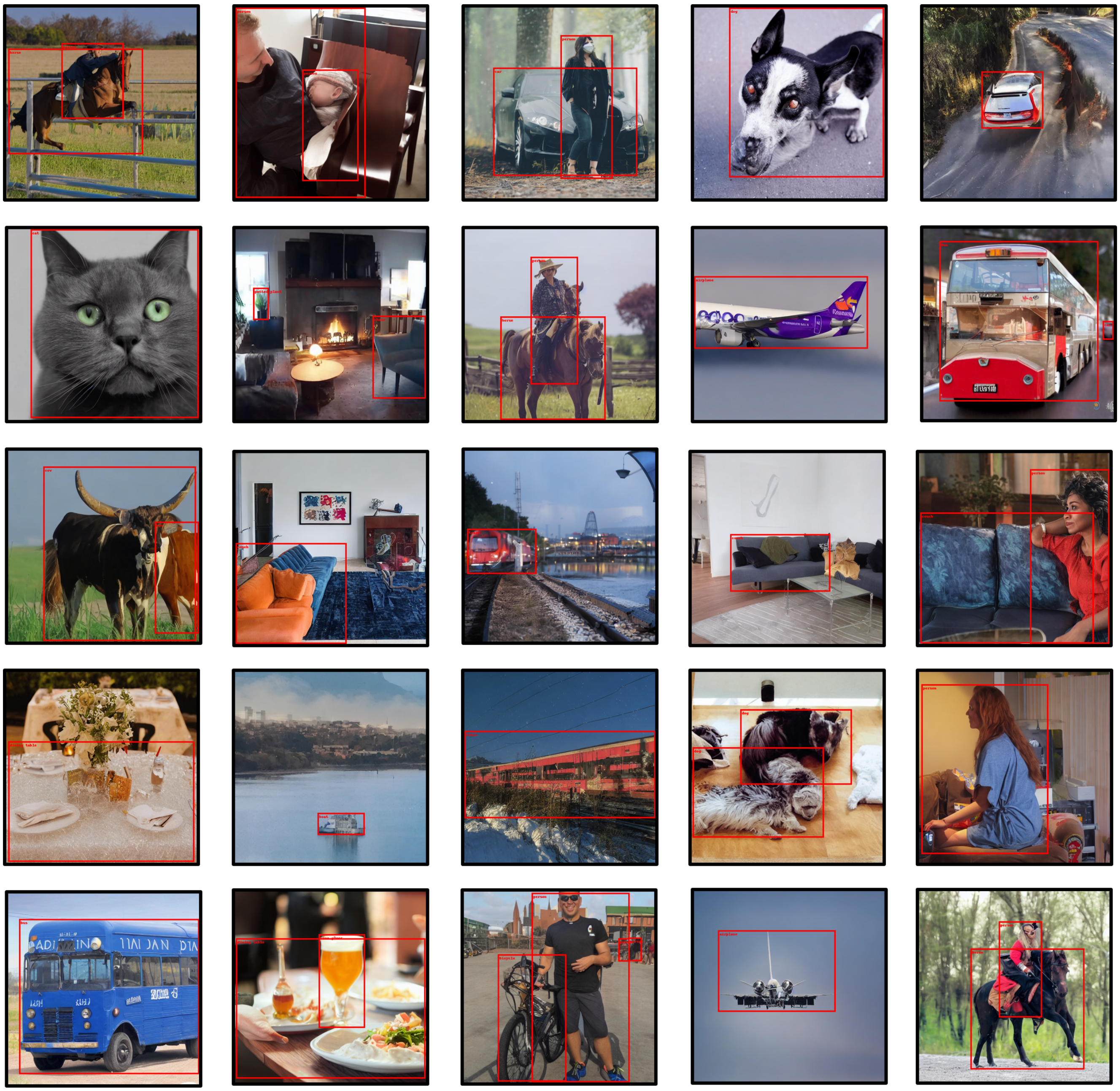

Figure 11. More visualizations selected from the distilled Pascal VOC on detection task.

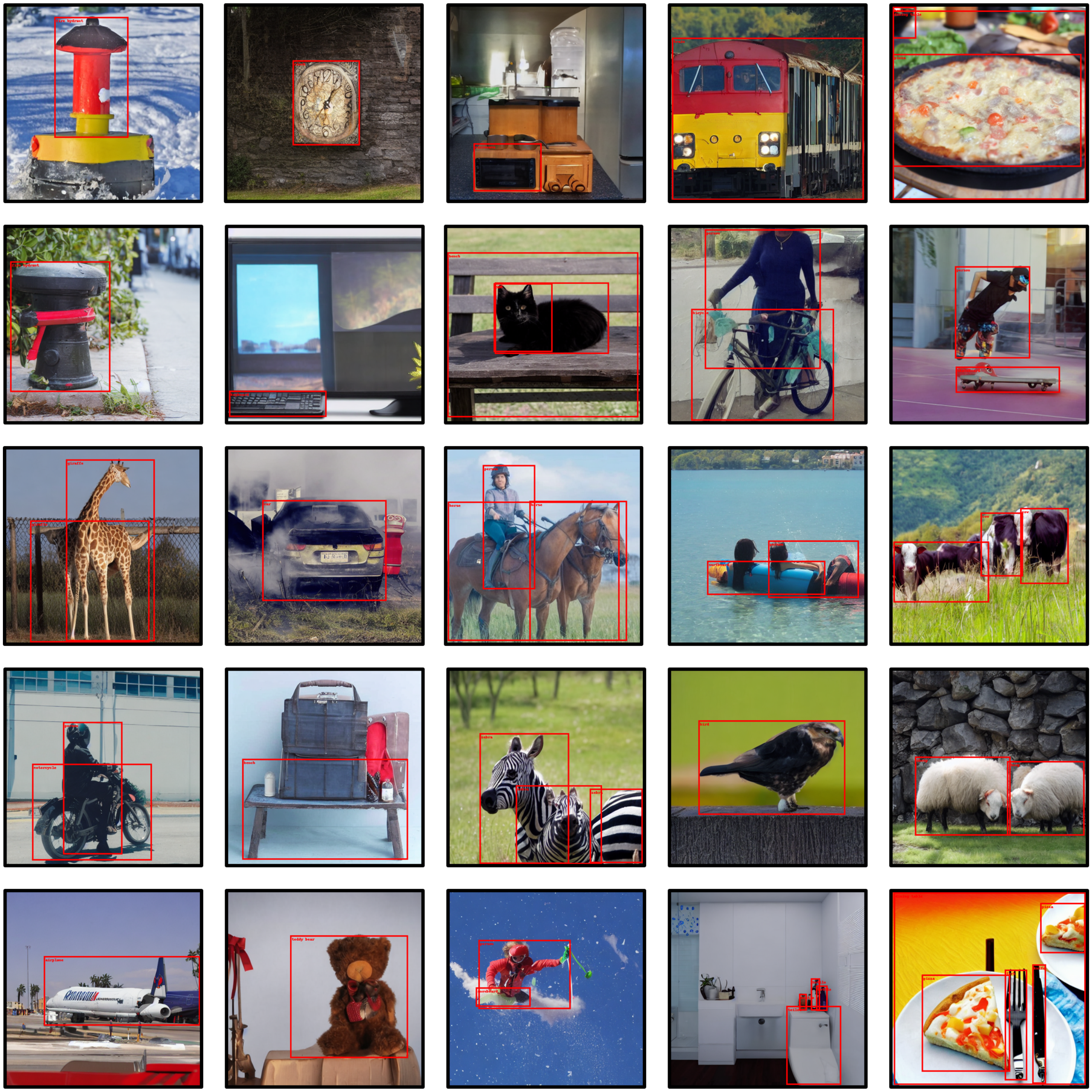

Figure 12. More visualizations selected from the distilled MS COCO on detection task.

CVPR
#360

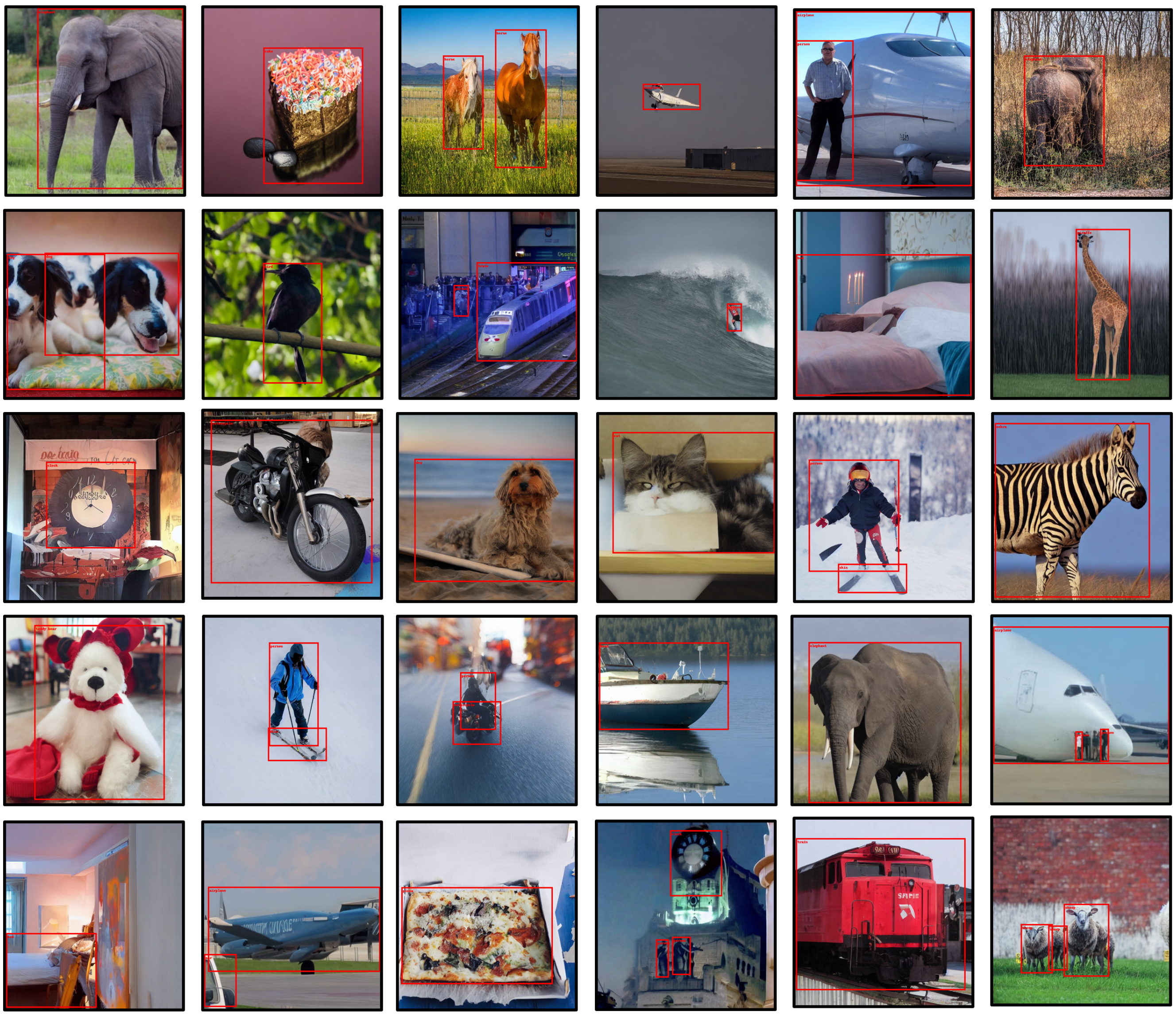

Figure 13. More visualizations selected from the distilled MS COCO on detection task.

CVPR
#360

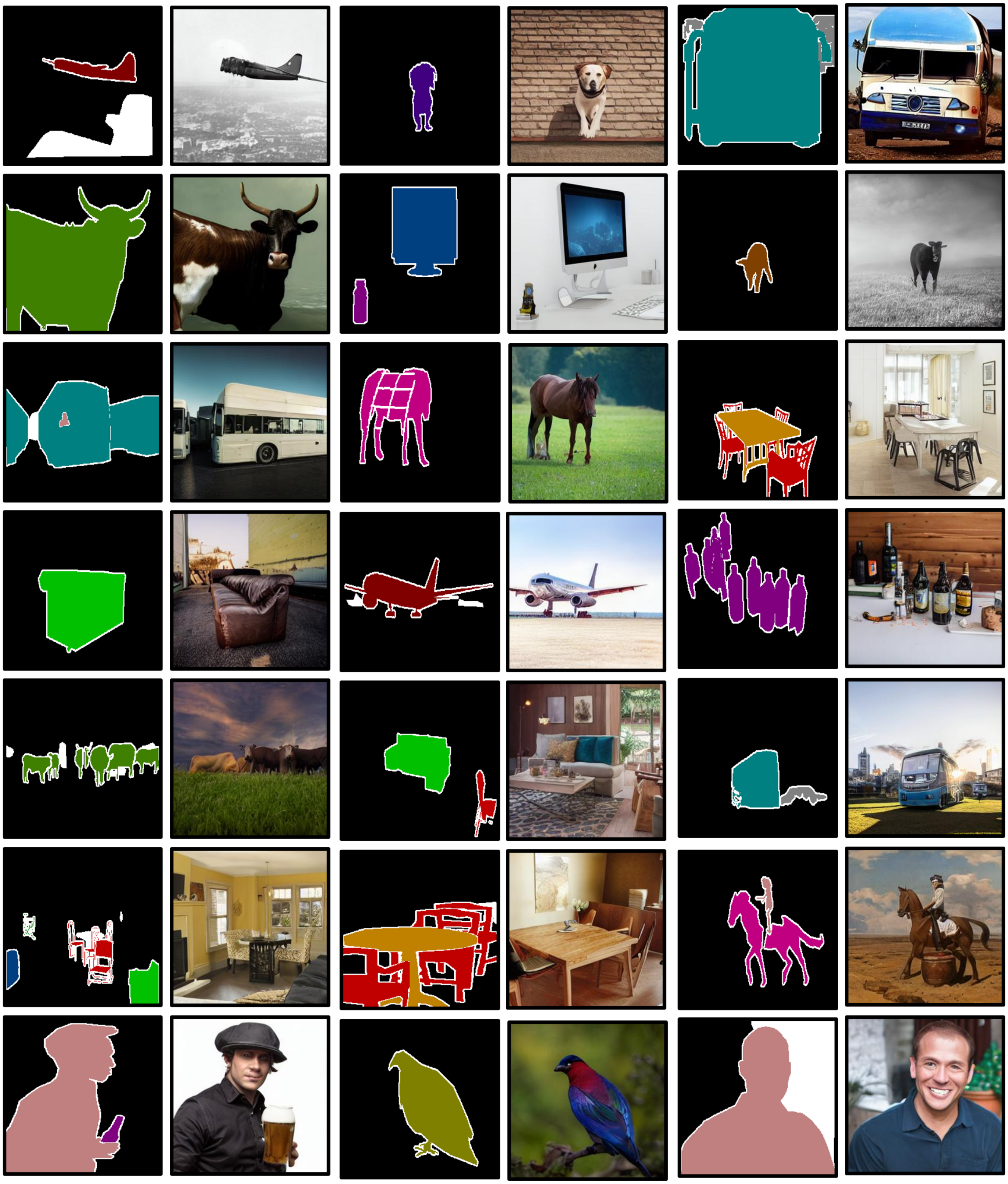

Figure 14. More visualizations selected from the distilled Pascal VOC on segmentation task.