Supplementary Materials to "Dataset Distillation via Vision-Language Category Prototype"

The supplementary material is organized as follows: Section 1 provides detailed related works; Section 2 presents the implementation details of generating textual descriptions using the LLaVA model used in this paper; Section 3 shows the details of diffusion model training; Section 4 presents implementation details of outlier removal and image prototypes; Section 5 presents more analysis and discussion; Section 6 provides detailed settings of hyperparameters; and finally, Section 7 presents generated samples of different datasets.

1. Related Works

1.1. Dataset Distillation

DD aims to synthesize information-dense, small datasets that effectively serve as alternatives to large-scale datasets for downstream tasks such as classification [6, 23], achieving performance comparable to that of the original data [20, 24]. Previous methods for DD primarily include meta-learning based and matching based frameworks [8]. In meta-learning based method, distilled data are optimized as hyperparameters within a bi-level framework, where synthetic data are iteratively updated in an outer loop by minimizing the meta-test loss on the real dataset, while an inner loop concurrently trains the model on the synthetic data [14, 20, 28]. Meta-learning methods are categorized into two sub-categories based on inner-loop optimization: backpropagation through time [2, 17, 20] and kernel ridge regression [13, 14] approach.

In contrast to the previously outlined meta-learning method, matching based methods refine distilled images via parameter matching and distribution matching [3]. Specifically, parameter matching focuses on aligning model parameters of the original dataset and synthetic dataset, both of which are trained based on the same network architecture. Zhao et al. [25] first proposed this approach to distilled images by matching the gradients computed on the synthetic dataset and real data, which is further developed through the following studies [1, 7]. Unlike encouraging the consistency of trained neural parameters, distribution matching approaches directly align the distribution of synthetic and real images in the feature space to closely approximate the real data. For instance, distribution matching minimizes the divergence between the two distributions by employing Maximum Mean Discrepancy (MMD) [27] metrics and is further extended by CAFE [19].

Though the methods mentioned above have achieved significant progress in dataset distillation, generating synthetic datasets often requires high computational costs, and the time involved increases rapidly with large-scale datasets. For instance, MTT distills images with IPC 50 on CIFAR10, requiring 47GB of GPU memory, making it impractical for large-scale datasets like ImageNet-1K [24]. Hence, several studies have leveraged generative models [4, 16] in dataset distillation to optimize latent features instead of image pixels, thereby achieving a faster training process and enhancing performance. For instance, IT-GAN [26] employs pretrained GAN as an informative training sample generator, which only needs to optimize latent features rather than image pixels. Su et al. [16] integrated the diffusion model into DD to obtain embedding features, after which K-means clustering is applied to each class. The cluster centers then serve as prototypes and, along with label texts, are fed into the diffusion model for image generation. However, existing diffusion models for dataset distillation often produce unrealistic images or images with absent target objects and exhibit co-occurrence bias. Our study overcomes these challenges by integrating text descriptions, enhancing logical coherence, and mitigating bias.

1.2. Vision-Language

Vision-Language (VL) is a multimodal learning approach that enables models to process and understand both visual (e.g., images and videos) and language (e.g., text and speech) information simultaneously, allowing for cross-modal understanding and reasoning beyond traditional unimodal models. Most multimodal methods project images and text into a shared embedding space, enabling the measurement of similarities between learned representations across diverse modalities, such as CLIP [15], BLIP [9] ALIGN [5]. These methods are foundational in enabling robust image-text alignment. Recent studies have extended these methods to multimodal dataset distillation. For instance, Wu et al. [21] leverage low-rank adaptation to match bi-trajectory in complex modern vision-language models. Similarly, Xu et al. [22] distill a ground truth similarity matrix from image-text pairs and employ low-rank factorization to improve efficiency and scalability. However, the absence of paired text in most datasets makes it challenging to apply the aforementioned methods, as they require multimodal alignment to achieve effective distillation.

To address this limitation, our method generates paired text for unimodal data via LLaVA [10–12] and integrates text prototypes with image prototypes to facilitate dataset distillation. This approach addresses the limitation of missing text in unimodal datasets while enabling flexible and scalable dataset distillation. Unlike prior methods [21, 22], which rely on image-text paired data for multimodal alignment, our approach is specifically tailored for datasets without textual annotations. By generating paired text and introducing text prototypes, our method achieves alignment between visual and textual information at the prototype level, creating a unified representation that improves the effectiveness of the distillation process. This innovation demonstrates the high adaptability of our approach to a broader range of datasets compared to existing methods.

2. Implementation of Generating Textual Descriptions Using the LLaVA Model

In this study, we employ the LLaVA model (liuhaotian/llava-v1.5-7b) [10–12] to generate textual descriptions for visual inputs. The model is run under default configurations, including a sampling temperature of 0.2 and a beam search with a single beam (num_beams=1). The generated textual descriptions are systematically stored for downstream analysis. All datasets follow a standardized prompt format to guide the model in generating the textual descriptions:

Prompt = “Describe the physical appearance of the {$CLASSNAME} in the image. Include details about its shape, posture, color, and any distinct features.”

Below is a sample of an English Springer Spaniel, along with the corresponding description generated by the model:

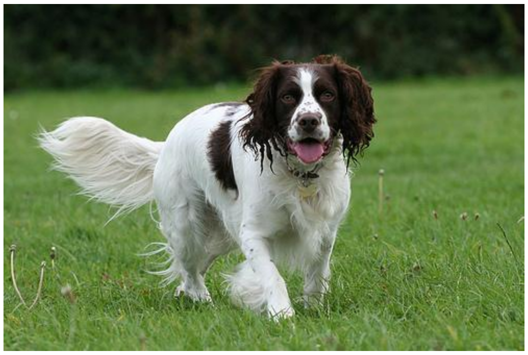


Figure 1. A sample of an English Springer Spaniel.

Description = “The English springer spaniel in the image is a small, white and brown dog with a short, curly coat. It has a distinctive shape, with a long body and a curved tail. The dog appears to be in a relaxed posture, walking through the grass with its tongue out, which is a common behavior for dogs when they are enjoying themselves or feeling comfortable. The dog’s coloration is predominantly white with brown markings, which is a common characteristic of English springer spaniels.”

The generated textual descriptions adhere to the standardized prompt, providing detailed accounts of the object’s shape, posture, color, and distinct features. These descriptions go beyond the visual content of the image, offering additional information not immediately apparent. By providing such comprehensive and nuanced details, the textual descriptions contribute valuable insights that cannot be captured by conventional image prototypes, thereby enriching the depth of the analysis.

3. Diffusion Model Training

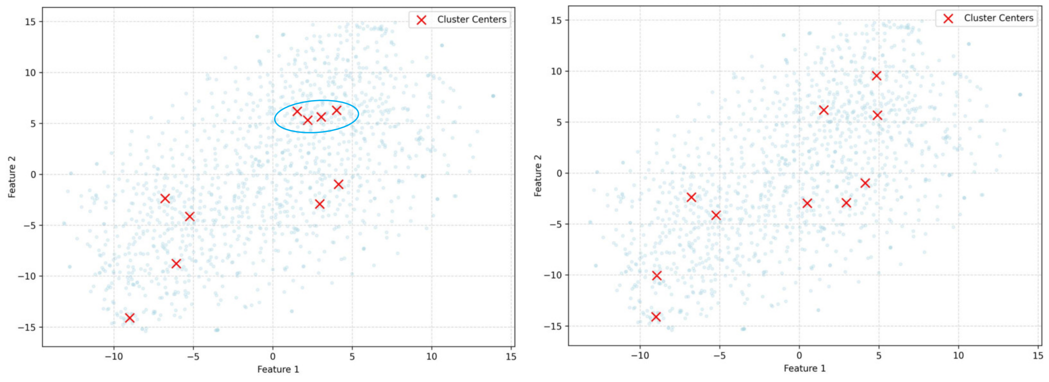
In this study, we fine-tune the benjamin-paine/stable-diffusion-v1-5 model using paired image-text datasets generated by LLaVA. To ensure robustness and model generalization, each dataset is trained with three distinct random seeds, resulting in three independently fine-tuned diffusion models per dataset. All datasets are trained according to the parameters specified in Table 1, except for the input resolution: 512 for ImageNet-1K and its subsets (ImageWoof, ImageNette, ImageIDC, ImageNet-100), 32 for CIFAR-10/100, and 64 for Tiny-ImageNet.

Settings	Values
train batch size	8
validation epochs	2
number of training epochs	8
learning rate	1.0e-05
Adam weight decay	0.01
augmentation	center crop, random flip

Table 1. Diffusion model training settings.

4. Outlier Removal and Image Prototypes

In contrast to other outlier detection methods, the Local Outlier Factor (LOF) is an unsupervised algorithm that does not require labeled data, making it particularly suitable for detecting anomalies in datasets with complex or unknown structures. LOF assesses the local density of data points to identify instances that significantly deviate from their surrounding neighbors. For the LOF algorithm, two parameters are configured: the number of neighbors



(a)

(b)

Figure 2. T-SNE visualizations of the bonnet class from ImageIDC. (a) Cluster centers obtained from K-means clustering on the original data. (b) Cluster centers after applying Local Outlier Factor (LOF) to remove outliers, followed by K-means clustering.

($n\_neighbors = 10$) is consistently applied across all datasets to maintain uniformity in local density estimation, while the contamination parameter is adjusted according to the specific characteristics of each dataset—set to 0.0 for ImageWoof, 0.1 for ImageIDC, 0.01 for ImageNette, and 0.1 for ImageNet-100.

As shown in Fig. 2, outlier removal improves clustering performance. In the original dataset (Fig. 2-(a)), K-means clustering resulted in several cluster centers being densely concentrated in specific regions, indicating poor cluster separation and potential overlap among clusters. This outcome suggests that the presence of outliers adversely affected the clustering process, leading to suboptimal partitioning of the data. However, after applying LOF to eliminate outliers (Fig. 2-(b)), the cluster centers appeared more distinctly separated and uniformly distributed across the feature space. This clearer separation reflects a more accurate representation of the underlying data structure, allowing K-means to perform more effectively.

Following outlier removal, K-means clustering is applied to each class individually, with the resulting cluster centers designated as image prototypes. The clustering process is configured with $random\_state$ set to the corresponding seed for reproducibility, $n\_init = 10$ to ensure stability in clustering outcomes, and the number of clusters ($n\_clusters$) determined by the images-per-class (IPC) setting. Specifically, when IPC = 10, the number of clusters is set to 10. This approach ensures that the prototypes effectively capture the representative features of each class after noise reduction.

5. More Analysis and Discussion

ImageNet-100 Beyond the previously mentioned 10-class ImageNet subsets, we also conduct experiments on the more challenging ImageNet-100, with the results presented in Table 2. This table provides a detailed performance analysis of state-of-the-art methods, including Random, Herding, IDC-1, Minimax, D4M, and our proposed approach (Ours), across various model architectures and IPC settings. To ensure a fair comparison across methods, all images are resized to 256 × 256. Although IDC-1 achieves the best performance at IPC = 10, our method outperforms IDC-1 at IPC = 20. Moreover, our approach consistently surpasses Minimax under both IPC = 10 and IPC = 20 settings.

IPC	Test Model	Random	Herding	IDC-1	Minimax	D4M	Ours
10	ConvNet-6	$17.0_{\pm0.3}$	$17.2_{\pm0.3}$	$\mathbf{24.3_{\pm0.5}}$	$20.1_{\pm0.3}$	$18.5_{\pm0.8}$	$\underline{22.3_{\pm0.2}}$
	ResNetAP-10	$19.1_{\pm0.4}$	$19.8_{\pm0.3}$	$\mathbf{25.7_{\pm0.1}}$	$21.5_{\pm0.3}$	$20.0_{\pm0.4}$	$\underline{24.5_{\pm0.1}}$
	ResNet-18	$17.5_{\pm0.5}$	$16.1_{\pm0.2}$	$\mathbf{25.1_{\pm0.2}}$	$20.1_{\pm0.9}$	$18.5_{\pm1.2}$	$\underline{23.3_{\pm0.5}}$
20	ConvNet-6	$24.8_{\pm0.2}$	$24.3_{\pm0.4}$	$\underline{28.8_{\pm0.3}}$	$25.9_{\pm0.4}$	$25.0_{\pm0.6}$	$\mathbf{29.3_{\pm0.6}}$
	ResNetAP-10	$26.7_{\pm0.5}$	$27.6_{\pm0.1}$	$\underline{29.9_{\pm0.2}}$	$27.0_{\pm0.4}$	$27.3_{\pm0.6}$	$\mathbf{32.3_{\pm0.6}}$
	ResNet-18	$25.5_{\pm0.3}$	$24.7_{\pm0.1}$	$\underline{30.2_{\pm0.2}}$	$26.4_{\pm0.3}$	$26.6_{\pm0.6}$	$\mathbf{31.5_{\pm1.1}}$

Table 2. Comparison of state-of-the-art methods on ImageNet-100 under various IPC settings and model architectures. The best results are marked as bold, and the second-best are underlined.

Tiny-ImageNet We evaluate our method on the Tiny-ImageNet dataset consisting of 200 classes, each containing 500 training images at 64 × 64 pixels. Table 3 presents the performance of different methods under IPC values, following the validation protocol of RDED [18]. When IPC = 10, our method achieves an accuracy of 42.6%, slightly surpassing RDED (41.9%), while SRe2L exhibits significantly lower performance (16.1%). At IPC = 50, RDED reaches the highest accuracy (58.2%), whereas our method achieves 55.5%, outperforming SRe2L (41.1%). These results indicate that our approach is competitive in low IPC but has room for improvement as IPC increases.

6. Hyper-Parameters Settings

In addition to the parameters α (Contamination), β (Non-representative Threshold), and k (Top$k$ words) used in the generation of text prototypes, the diffusion model with text

Dataset	IPC	SRe2L	RDED	Ours
Tiny-ImageNet	10	16.1±0.2	41.9±0.2	**42.6±0.2**
	50	41.1±0.4	**58.2±0.1**	55.5±0.6

Table 3. Comparison of state-of-the-art methods on Tiny-ImageNet under various IPC settings.

prompts involves two hyperparameters: *strength* ($0 < s < 1$) and *guidance scale* ($g > 1$). *Strength* controls the degree of modification applied to the input image based on the text prompt, with lower values resulting in minimal modification, preserving the original image, and higher values leading to a stronger alignment with the prompt. *Guidance scale* determines the extent to which the model adheres to the text prompt during image generation. Lower values encourage more diverse outputs, while higher values enforce stricter alignment with the prompt.

For a fair comparison, ImageWoof, ImageNette, ImageIDC, and ImageNet-100 are evaluated using the Minimax valuation method at a synthetic image resolution of 256×256, as shown in Table 4. Additionally, apart from ImageWoof and ImageNet-100, which is evaluated on three widely used network architectures (ConvNet-6, ResNetAP-10, and ResNet-18), the remaining datasets (ImageNette and ImageIDC) are assessed only on ResNetAP-10.

Settings	**Values**
contamination	0.0/0.01/0.1/0.1
nonrepresentative threshold	0.7
top$k$ words	30
guidance scale	10
strength	0.7
synthetic image size	256
valuation method	Minimax

Table 4. Parameter settings of ImageWoof, ImageNette, ImageNet-100, and ImageIDC.

For ImageNet-1K, hyperparameters are provided in Table 5. All synthetic images are resized to 224 × 224, following the validation protocol of RDED with the ResNet-18 architecture.

For CIFAR10/CIFAR100, hyperparameters are summarized in Table 6. All synthetic images are resized to 32 × 32, following the validation protocol of RDED with the modified ResNet-18 architecture.

For Tiny-ImageNet, hyperparameters are outlined in Table 7. All synthetic images are resized to 64 × 64, following the validation protocol of RDED with the modified ResNet-18 architecture.

Settings	**Values**
contamination	0.0
nonrepresentative threshold	0.7
top$k$ words	30
guidance scale	10
strength	0.7
synthetic image size	224
valuation method	RDED

Table 5. Parameter settings of ImageNet-1K.

Settings	**Values**
contamination	0.0
nonrepresentative threshold	0.8/0.2
top$k$ words	30
guidance scale	10
strength	0.7
synthetic image size	32
valuation method	RDED

Table 6. Parameter settings of CIFAR10/CIFAR100.

Settings	**Values**
contamination	0.0
nonrepresentative threshold	0.7
top$k$ words	30
guidance scale	10
strength	0.7
synthetic image size	64
valuation method	RDED

Table 7. Parameter settings of Tiny-ImageNet.

7. Generated Samples of Different Datasets

Additional visualizations of the distilled data generated by our method are presented in the following figures: CIFAR-10 (Fig. 3), CIFAR-100 (Fig. 4), Tiny-ImageNet (Figs. 5–6), and Imagenet-1k (Figs. 7–16).

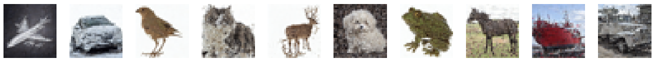


Figure 3. Synthetic images (32 × 32) selected from the distilled CIFAR-10 (Class 0-9).

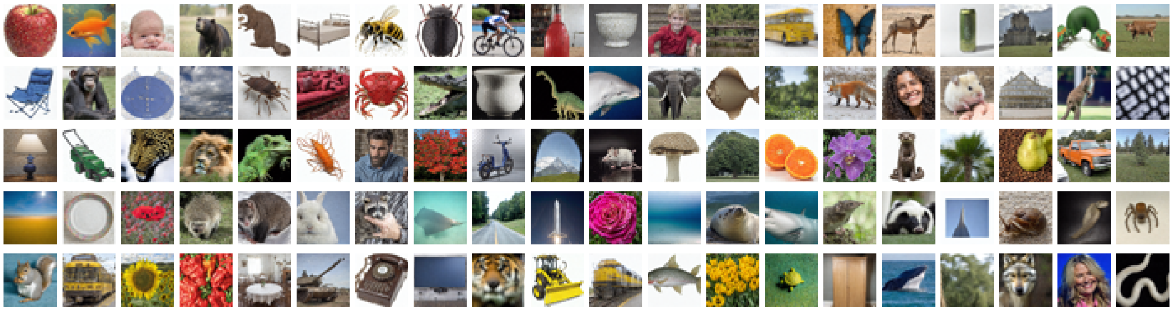


Figure 4. Synthetic images (32 × 32) selected from the distilled CIFAR-100 (Class 0-99)

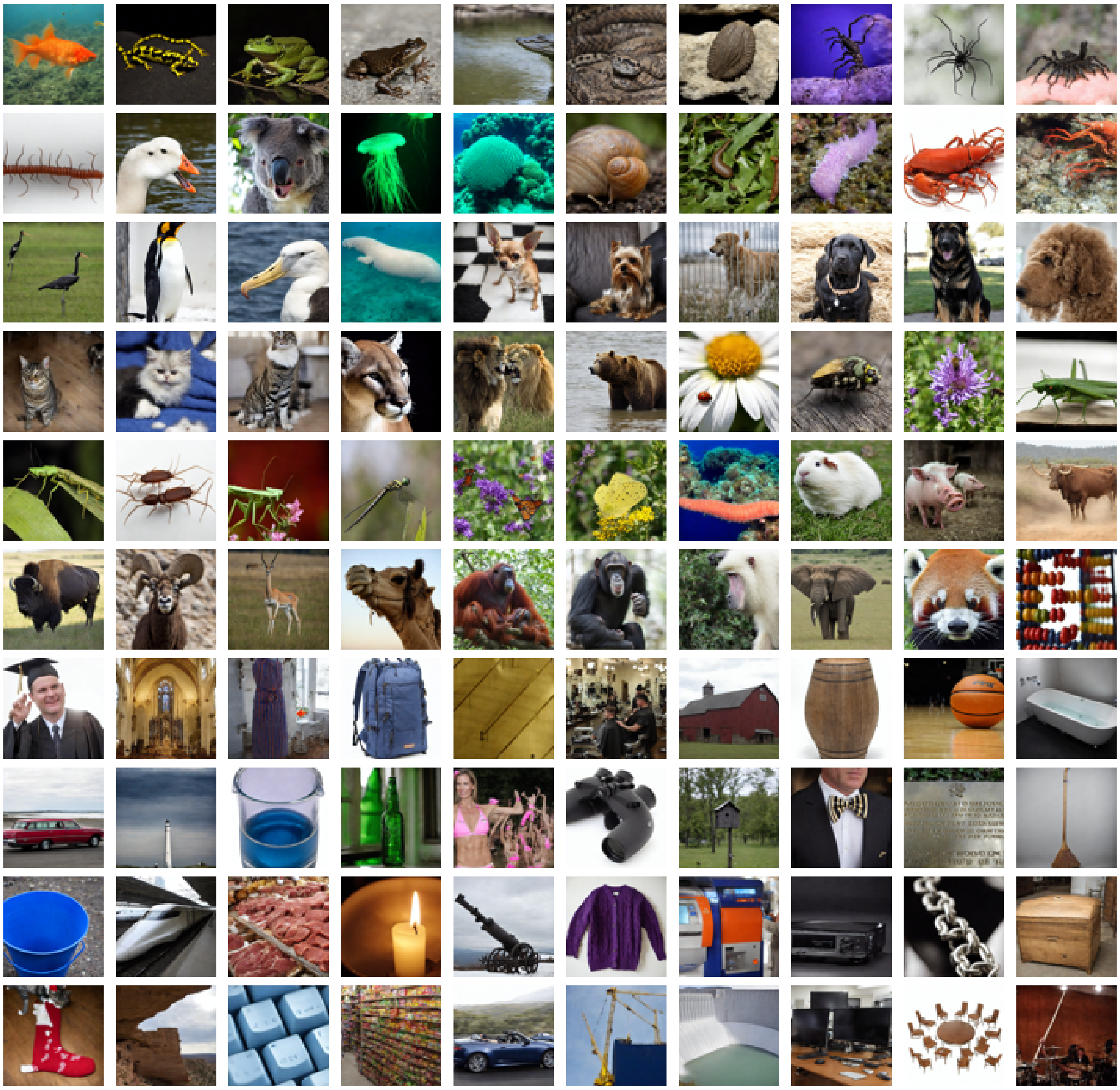


Figure 5. Synthetic images (64 × 64) selected from the distilled Tiny-ImageNet (Class 0-99).

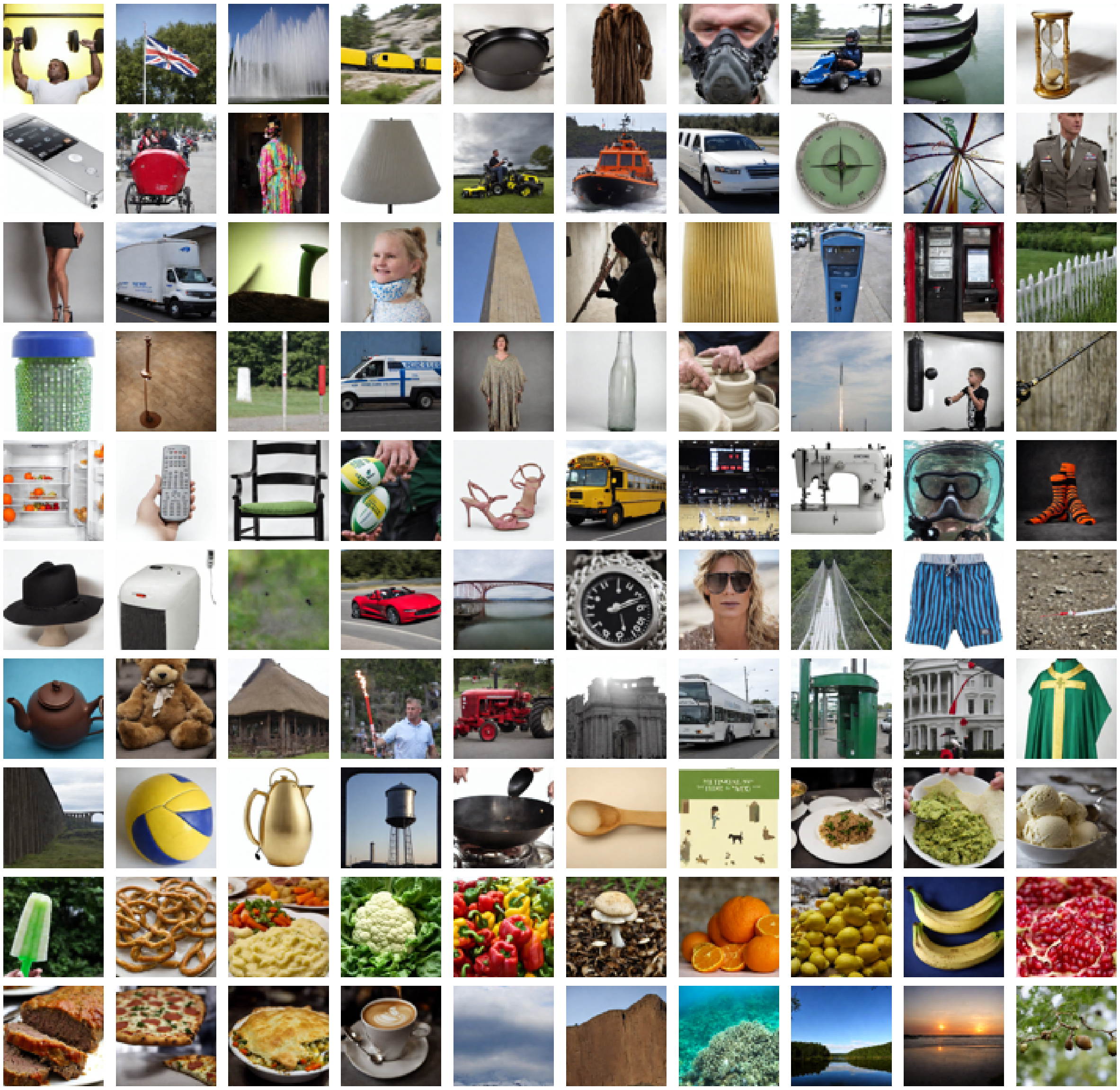


Figure 6. Synthetic images (64 × 64) selected from the distilled Tiny-ImageNet (Class 100-199).

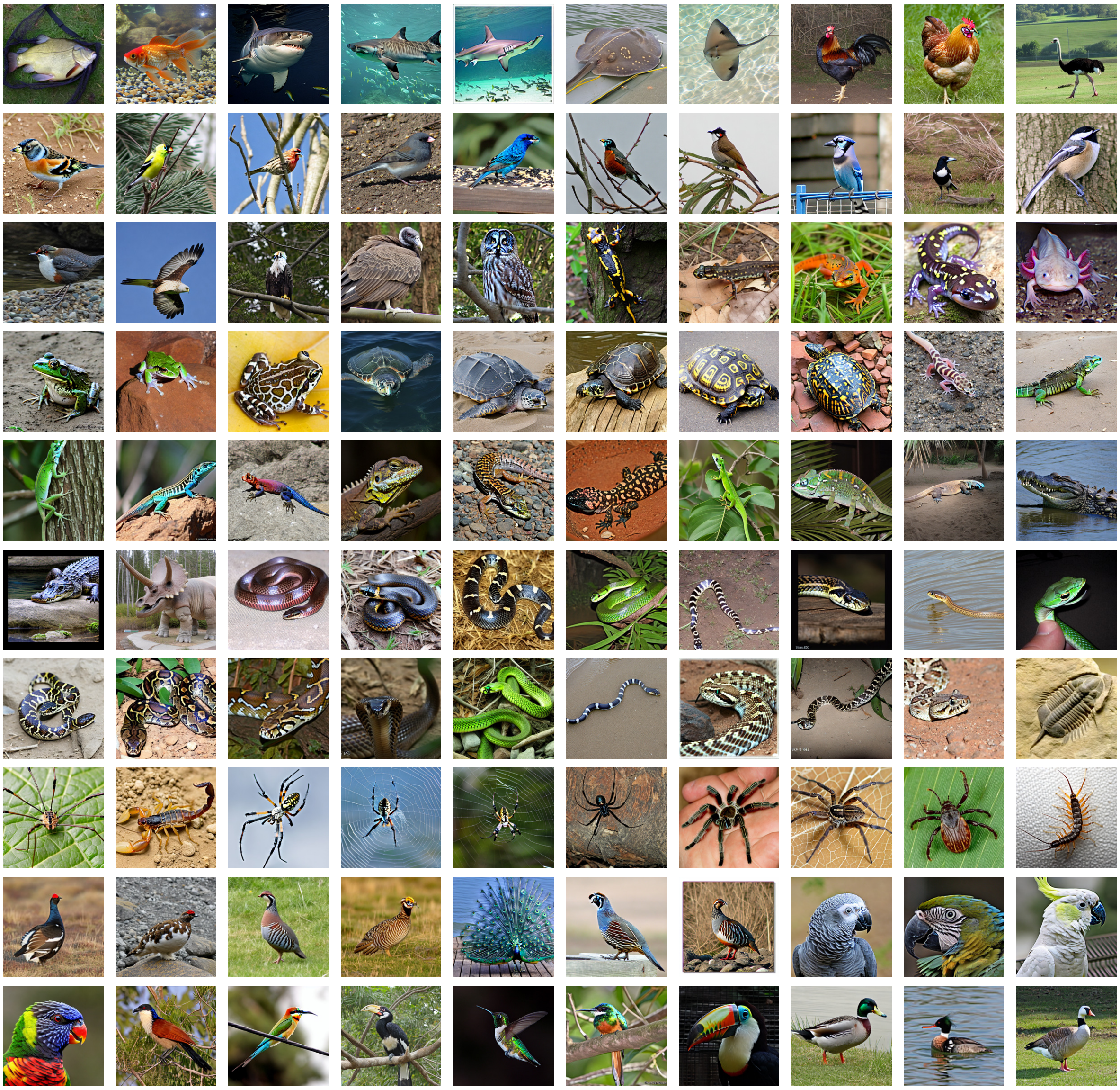


Figure 7. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 0-99).

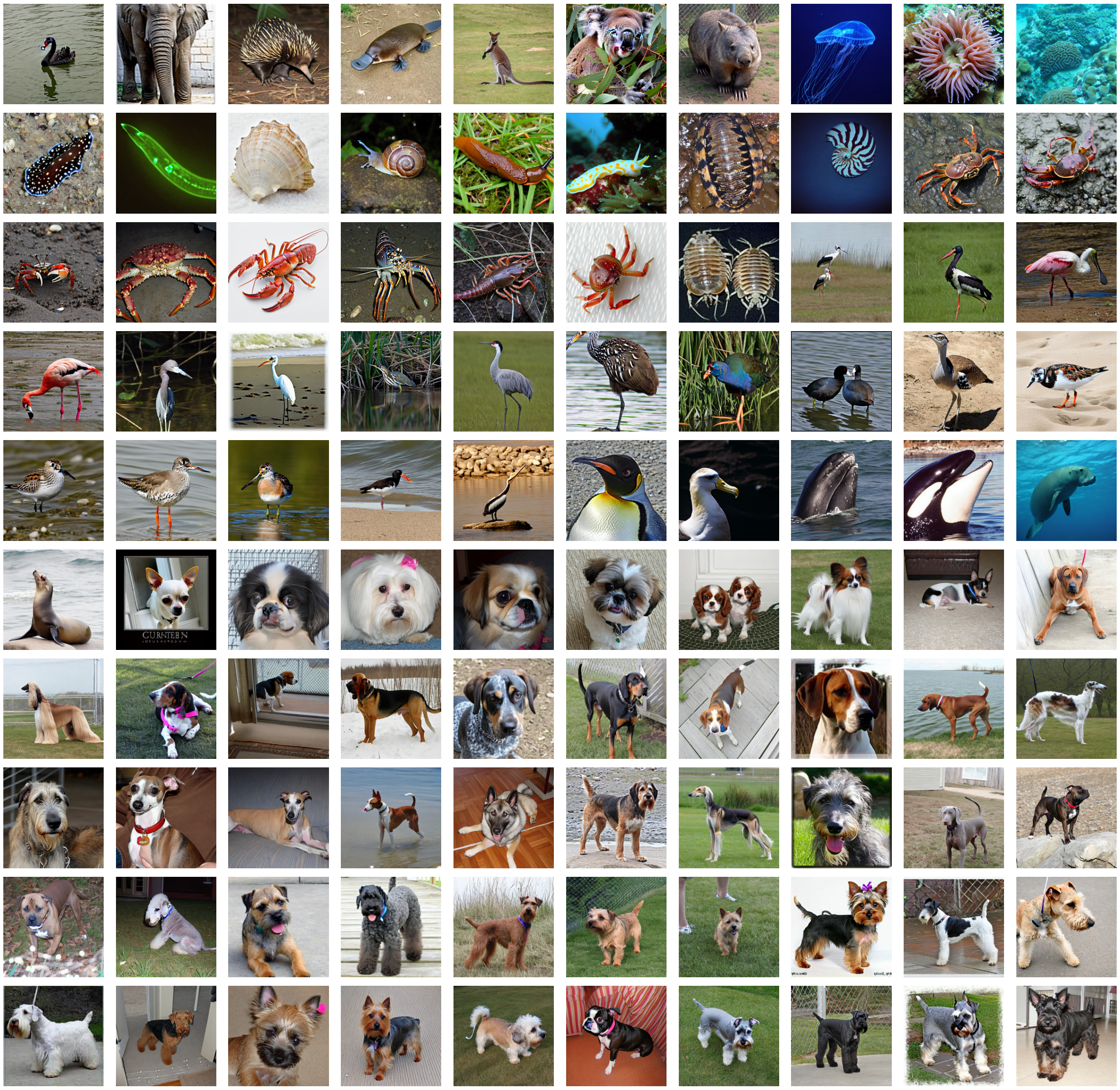


Figure 8. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 100-199).

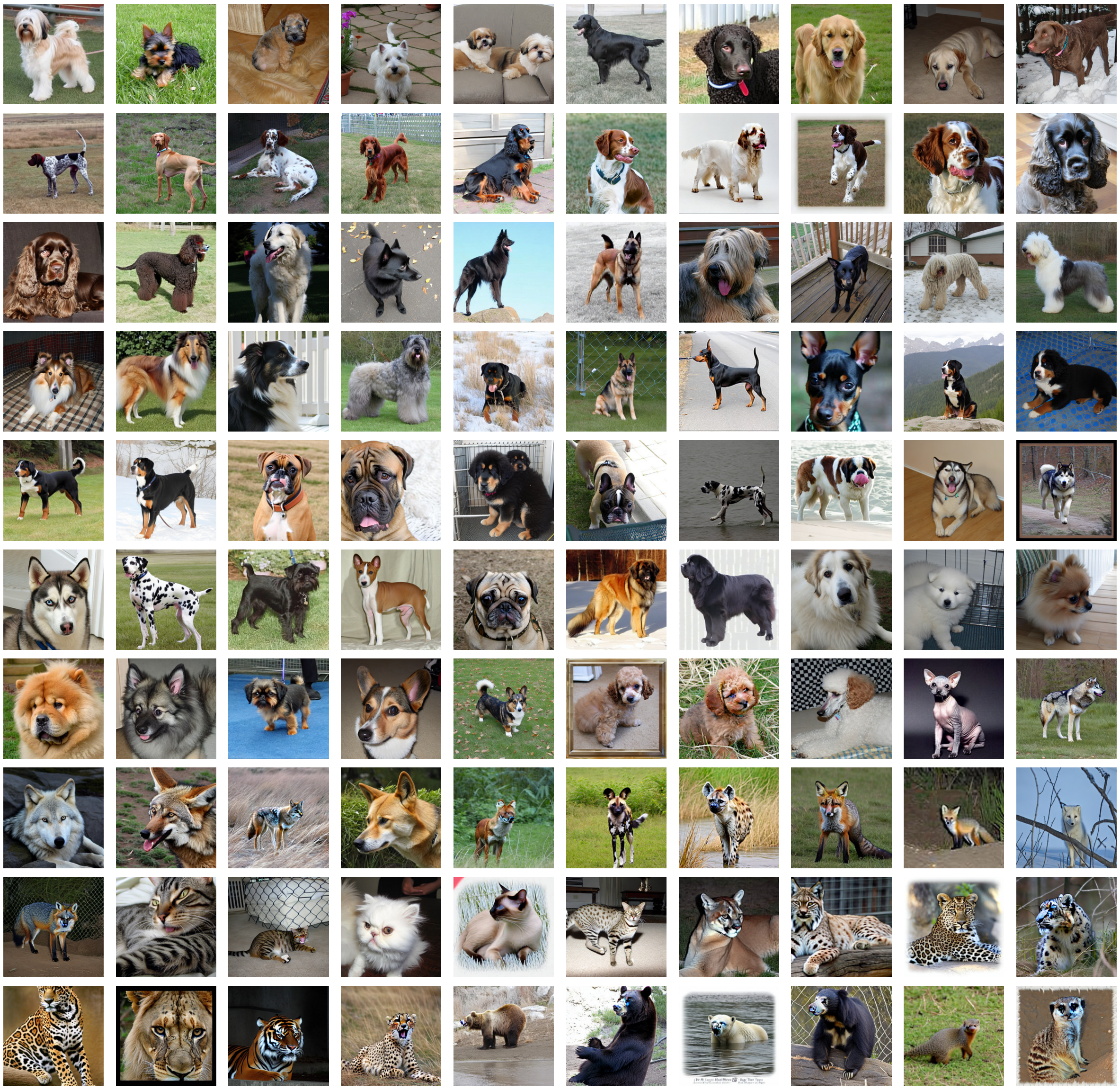


Figure 9. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 200-299).

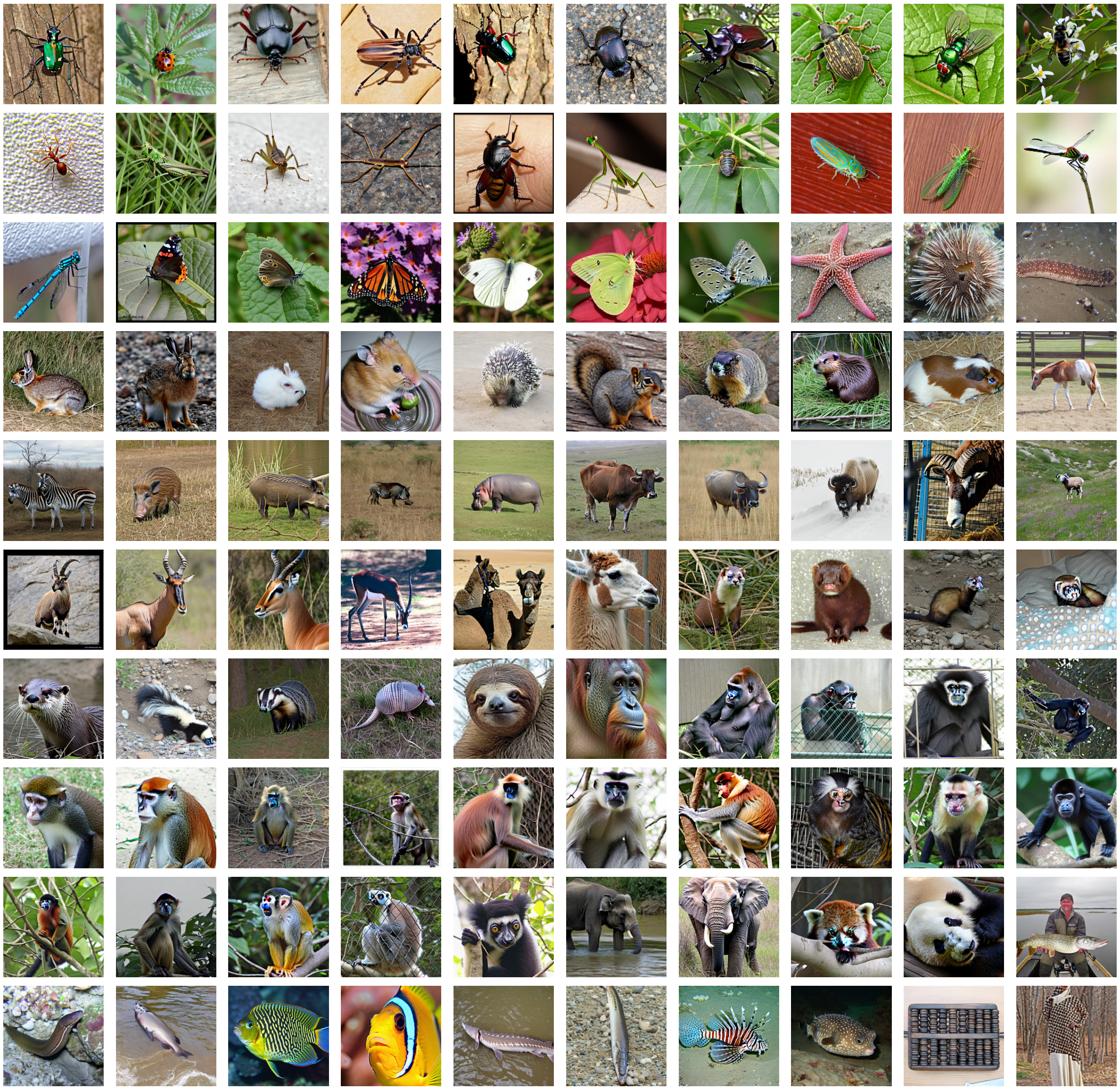


Figure 10. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 300-399).

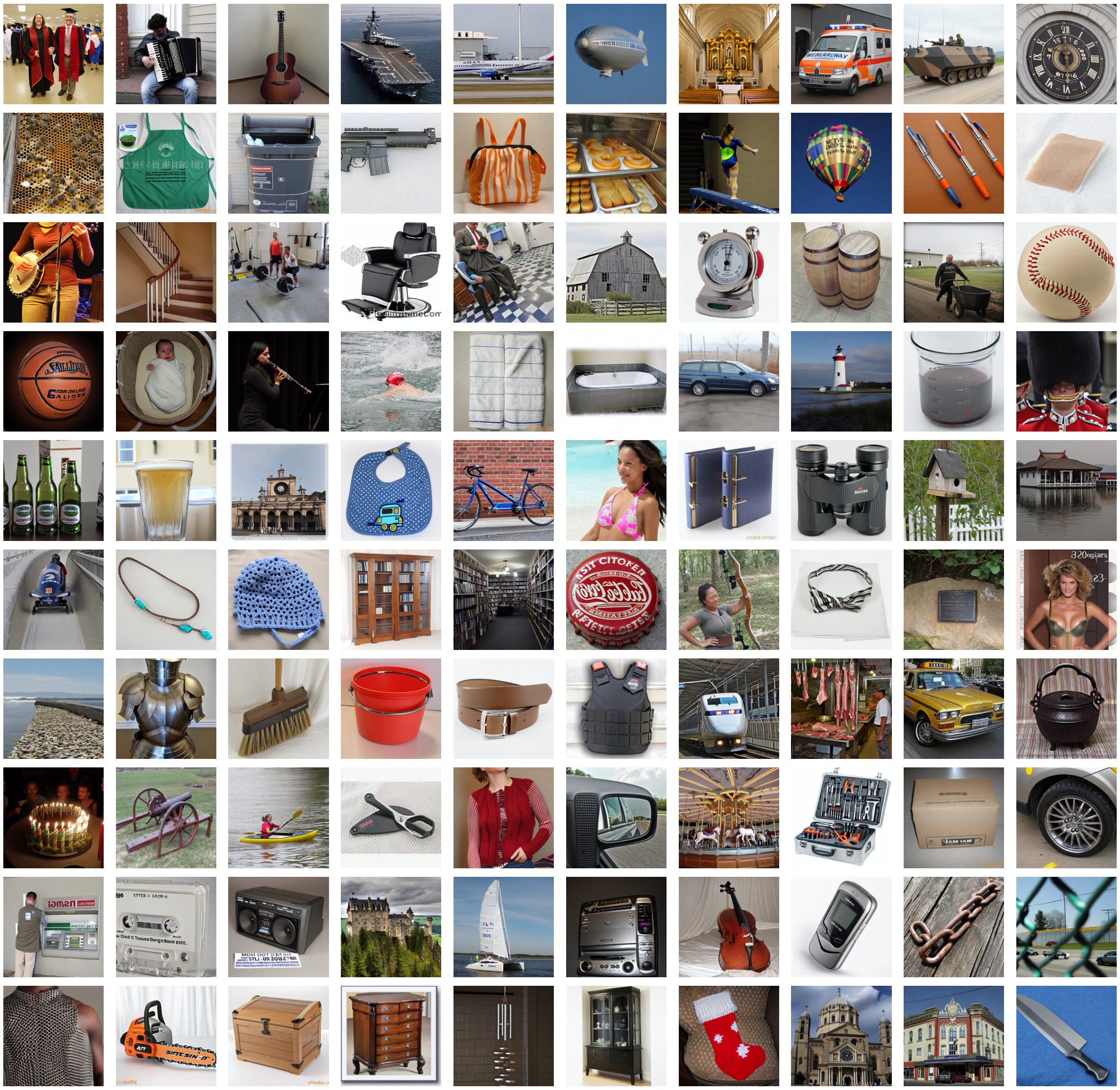


Figure 11. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 400-499).

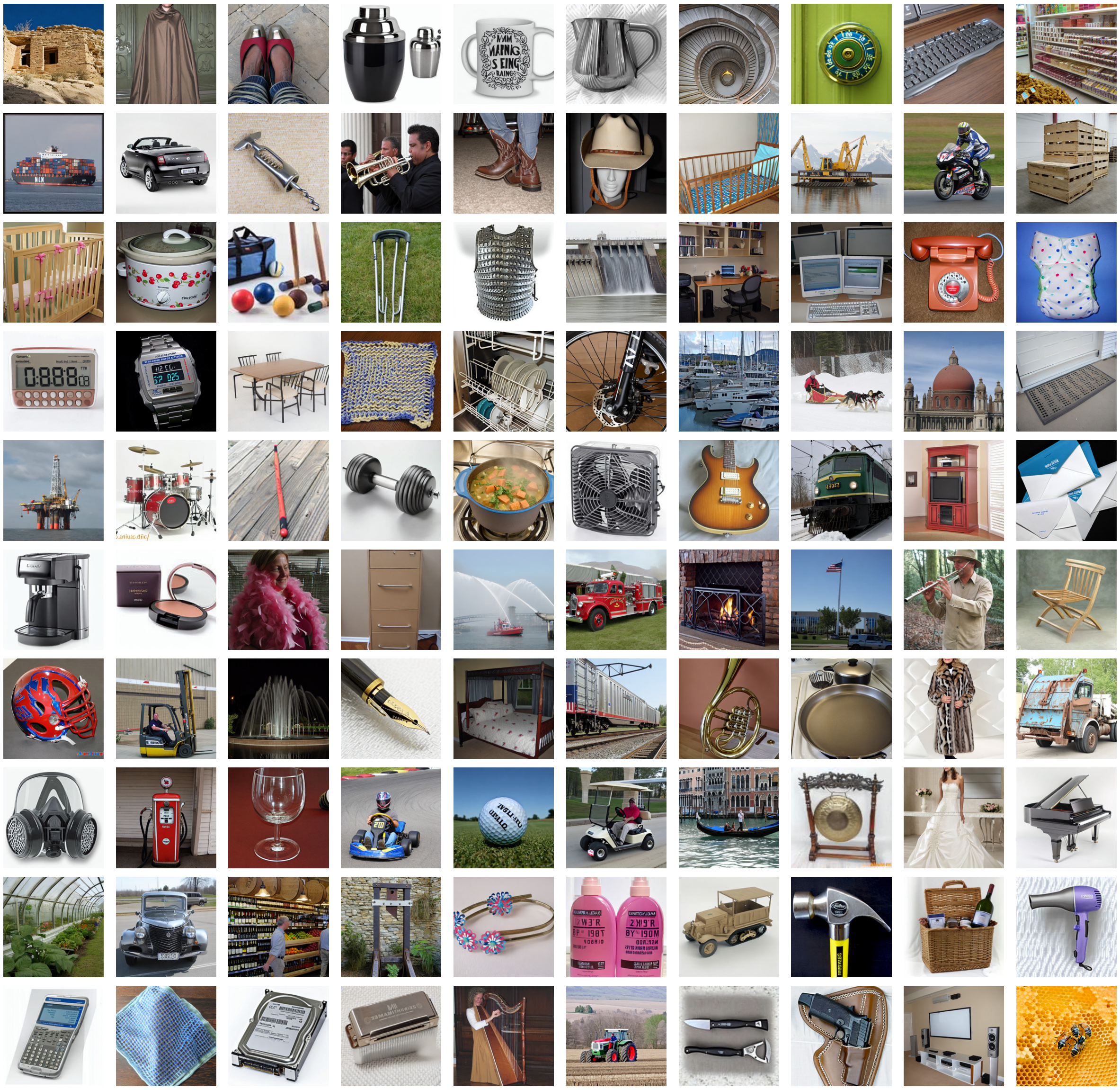


Figure 12. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 500-599).

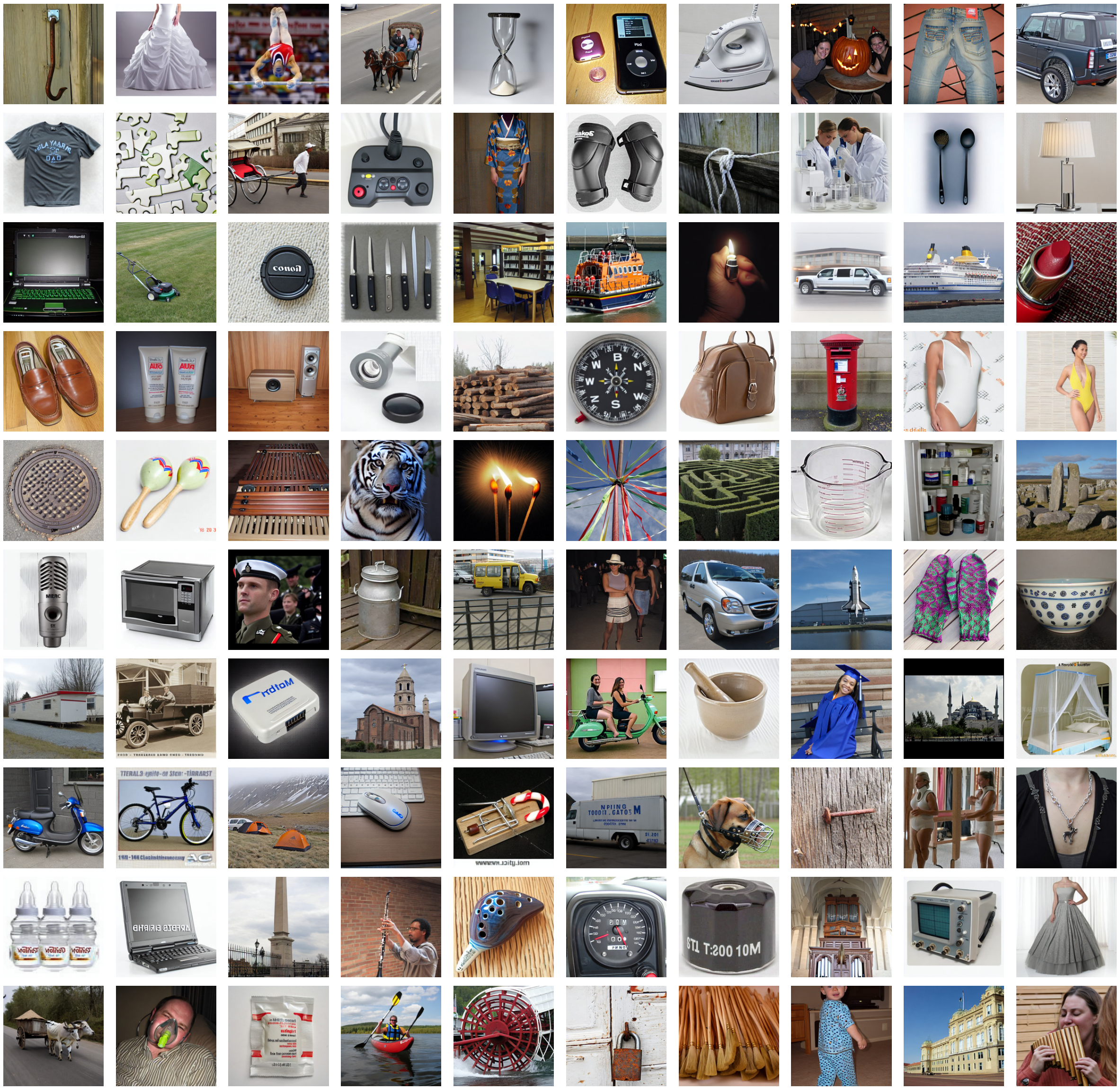


Figure 13. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 600-699).

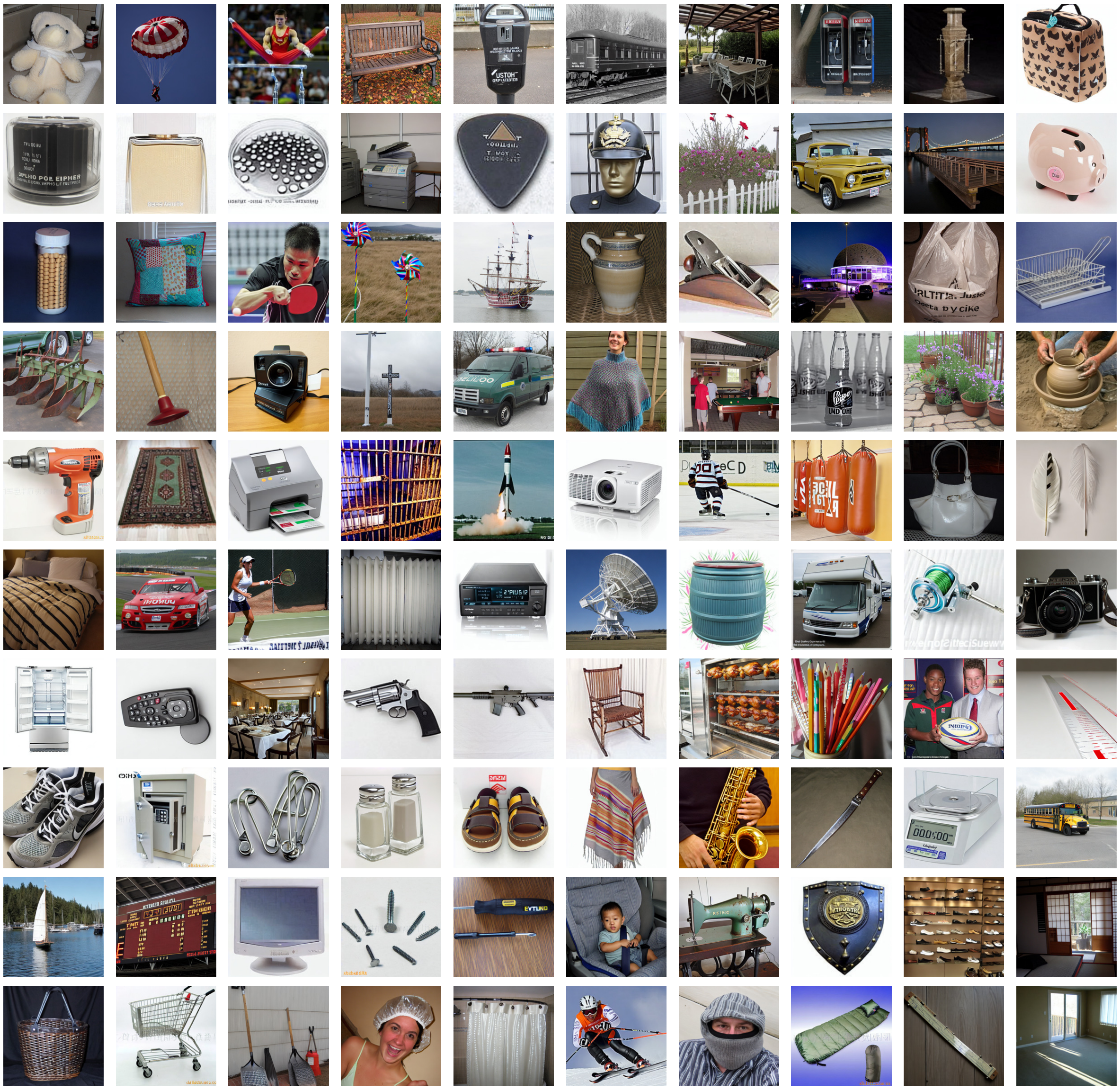


Figure 14. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 700-799).

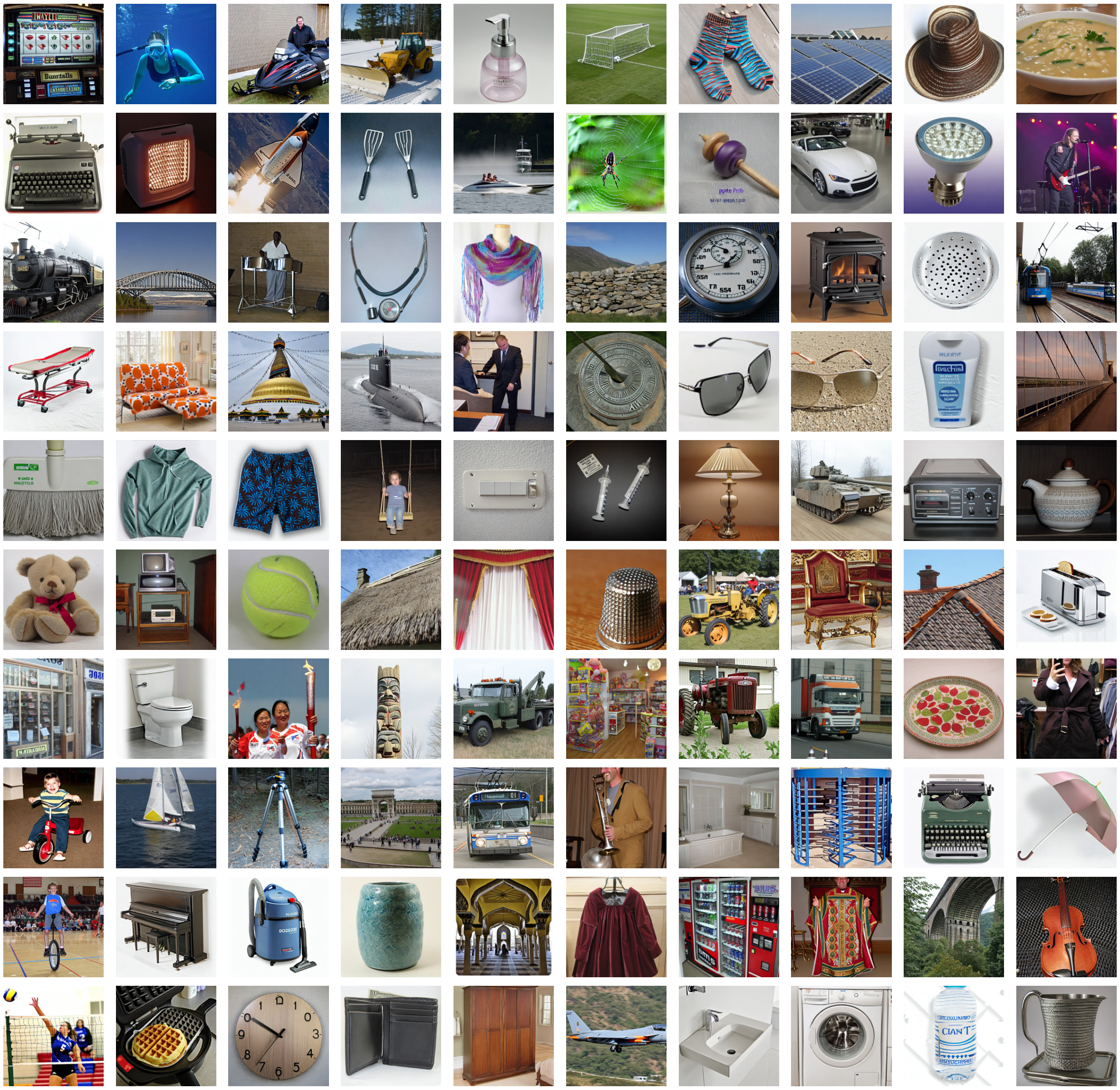


Figure 15. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 800-899).

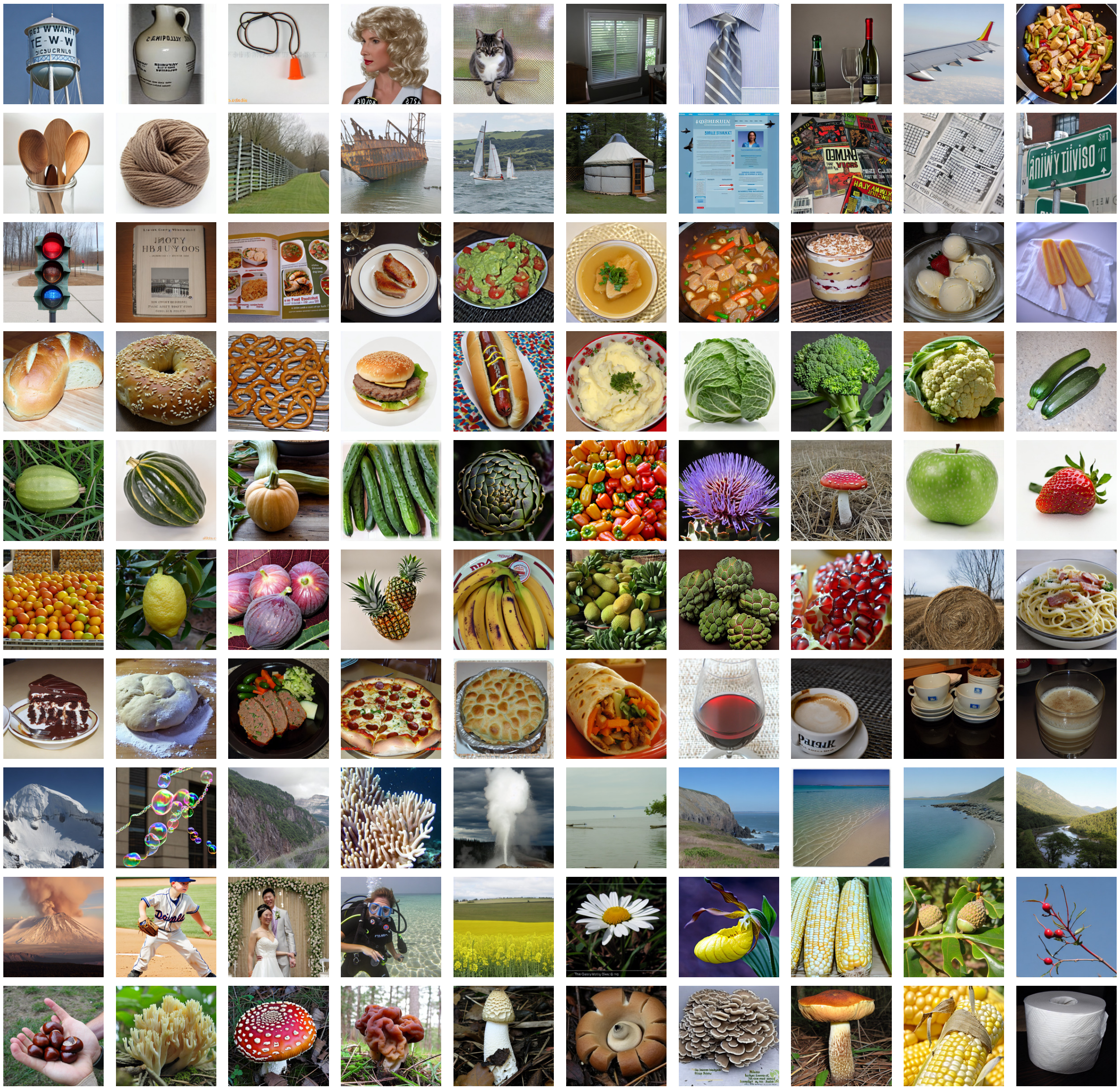


Figure 16. Synthetic images (224 × 224) selected from the distilled Imagenet-1k (Class 900-999).

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4750–4759, 2022. 1

[2] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 34391–34404, 2022. 1

[3] Zongxion Geng, Jiahui andg Chen, Yuandou Wang, Herbert Woisetschlaeger, Sonja Schimmler, Ruben Mayer, Zhiming Zhao, and Chunming Rong. A survey on dataset distillation: Approaches, applications and future directions. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2023. 1

[4] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15793–15803, 2024. 1

[5] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4904–4916, 2021. 1

[6] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 11102–11118, 2022. 1

[7] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12352–12364, 2022. 1

[8] Shiye Lei and Dacheng Tao. A comprehensive survey to dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):17–32, 2023. 1

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022. 1

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916, 2023. 2

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024. 2

[13] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 13877–13891, 2022. 1

[14] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1

[16] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. Dˆ 4: Dataset distillation via disentangled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5809–5818, 2024. 1

[17] Ilia Sucholutsky and Matthias Schonlau. Soft-label dataset distillation and text dataset distillation. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. 1

[18] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9390–9399, 2024. 3

[19] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12196–12205, 2022. 1

[20] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1

[21] Xindi Wu, Byron Zhang, Zhiwei Deng, and Olga Russakovsky. Vision-language dataset distillation. *Transactions on Machine Learning Research*, 2024. 1, 2

[22] Yue Xu, Zhilin Lin, Yusong Qiu, Cewu Lu, and Yong-Lu Li. Low-rank similarity mining for multimodal dataset distillation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 2

[23] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[24] Ruonan Yu, Songhua Liu, and Xinchao Wang. A comprehensive survey to dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):150–170, 2023. 1

[25] Bo Zhao and Hakan Bilen. Dataset condensation with gradient matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1

[26] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Workshop*, 2022. 1

[27] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6514–6523, 2023. 1

[28] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 9813–9827, 2022. 1