DiT-Distill: Open-Set Fine-Grained Retrieval via Generative Curriculum Knowledge

Supplementary Material

This supplementary material includes an introduction to Curriculum Learning (Sec. S1), more implementation details for constructing the CDR-DiT training data (Sec. S2), more experimental results and visualizations of DiT-Distill (Sec. S3), and a discussion of limitations (Sec. S4).

S1. Curriculum Learning

Curriculum learning is first introduced by Bengio et al. [3] as a training strategy that mimics the step-by-step learning process of humans. Later studies extend this idea. For example, SPL [20] and MCL [66] adjust data diversity using static sample difficulty, while DiH [67] adjusts the learning pace based on the exponential moving average of sample difficulty. Previous works [29, 42, 61] have also applied curriculum learning to more challenging domains such as noisy data environments and visual QA, demonstrating its potential in challenging scenarios. DynaCor [29] separates incorrectly labeled instances from correctly labeled ones through clustering of training dynamics. DCML [42] integrates curriculum learning with meta-learning to address noisy few-shot tasks. More recently, researchers have explored combining data augmentation with curriculum learning [1, 34, 38, 60]. CUDA [1] merges curriculum learning with image augmentation to improve performance on tail classes in long-tail learning. DisCL [34] adjusts the image-guidance level during image synthesis at each training stage to better learn from hard samples. In our work, we treat the generative knowledge at different diffusion timesteps as curriculum knowledge and transfer it step-by-step to the retrieval backbone.

S2. More Details for Constructing the CDR-DiT Training Data

This supplementary is for Sec. 3.2 of the main paper. Fig. S.2 illustrates the complete pipeline for constructing the training data for CDR-DiT, including generating *context-to-object* image pairs and their corresponding attribute-centric textual descriptions.

S2.1. Data Construction Details

Construction of Context-to-Object Image Pairs. The procedure for constructing context-to-object image pairs is summarized in Algorithm 1.

To isolate fine-grained objects from complex backgrounds, we employ an open-vocabulary detector [37]. Crucially, we use the **coarse super-category name** (*e.g.,* "bird" or "car") as the detection prompt P, rather than specific fine-grained labels. This ensures that our object extraction process remains generalizable and does not leak fine-grained category information, strictly adhering to the open-set setting.

Algorithm 1 The process of constructing context-to-image pairs

Input: Context image $\mathbf{I}$, text prompt P, open-vocabulary detector OVD
Output: Object Image $\mathbf{I}_O$

$score, \mathbf{I}_O \leftarrow \text{OVD}(\mathbf{I}, P)$
if $score < 0.5$ **then**
  $\mathbf{I}_O \leftarrow None$
end if

Table S.1. The number of context-to-object image pairs for each dataset.

Dataset	CUB-200-2011	Cars	Dogs	NABirds
Number	5,218	7,411	9,053	17,689

Given a context image $\mathbf{I}$ and prompt P, the detector extracts the object bounding box. We filter out low-confidence detections (score < 0.5) to ensure data quality. The resulting object image $\mathbf{I}_O$ effectively suppresses irrelevant background noise (see Fig. S.1).

Preprocessing Note: Since fine-grained recognition relies heavily on precise geometric cues (e.g., beak shape, wheel proportions), simply resizing $\mathbf{I}_O$ to a square resolution would introduce shape deformation. Therefore, we explicitly pad the shorter side of the cropped object with zeros (or mean pixel values) to verify the aspect ratio before resizing. The statistics of the constructed pairs are detailed in Table S.1.

Generation of Attribute-Centric Textual Descriptions. Using the clean object image $\mathbf{I}_O$, we generate the attribute-centric description $\mathbf{T}_{text}$ using Qwen2.5-VL-7B [2]. We designed a specific attribute-focused instruction: *"Describe the [cls] in the image and its characteristics in one sentence, without outputting its category name."* This negative constraint (*"without outputting..."*) is critical. It prevents the VLM from "taking a shortcut" by simply retrieving encyclopedic knowledge associated with a class name. Instead, it forces the VLM to ground its description in visual evi-

dence, resulting in rich descriptions of color, texture, and parts (as shown in Fig. S.1) that are essential for learning discriminative generative curriculum knowledge.

S3. More Experiments Analysis

S3.1. Ablation Study

This supplementary is for Sec. 4.3 of the main paper.

S3.1.1. Ablation on CDR Conditioning Components

A core premise of our method is that the CDR stage relies on **attribute-centric text guidance** ($\mathbf{T}_{text}$) to refine the DiT's focus. To validate this, we perform a comprehensive ablation study in Table S.2 by progressively degrading the conditioning signals.

Impact of Textual Guidance (A1 & A2). We first test the necessity of the attribute description.

- **A2 (No Text):** When we remove the text entirely (using an empty string), the model degrades to a pure image-inpainting/cropping task. Performance drops significantly from 87.2% to 85.3%. This -1.9% drop provides direct evidence that the visual context alone is insufficient for learning fine-grained discrepancies.
- **A1 (Generic Text):** Using a fixed template (*"A photo of a [cls]"*) recovers some performance (86.1%), likely by activating the DiT's general class priors.
- **Full Model (Attribute Text):** However, our full model with specific attribute descriptions achieves the best result (87.2%). The clear hierarchy (*No Text* $<$ *Generic* $<$ *Attribute*) strongly validates our "attribute-centric" hypothesis: it is the *specific semantic details* in $\mathbf{T}_{text}$ that guide the model to focus on critical visual discrepancies.

Impact of Contextual Guidance (A3). Finally, we investigate the role of the context image $\mathbf{I}$ (A3). Removing $\mathbf{I}$ reduces the task to standard text-to-image fine-tuning (generating $\mathbf{I}_O$ from text only). The performance drop to 86.4% confirms that the **context-to-object** formulation is essential. The context image acts as a "control signal", forcing the model to learn the discrepancy between the holistic context and the fine-grained object, rather than just generating the object from scratch.

S3.1.2. Impact of MLLM Selection

The quality of the text guidance $\mathbf{T}_{text}$ depends heavily on the Multimodal Large Language Model (MLLM) used to generate it. In Table S.3, we compare descriptions generated by three state-of-the-art MLLMs. Interestingly, we find that performance is not correlated with model size. The largest model, GLM-V4.1 (9B), achieves 86.9% R@1. However, the smaller Qwen2.5-VL (7B) outperforms it with 87.2% R@1. This counter-intuitive result suggests that for fine-grained understanding, the quality of visual-semantic alignment (i.e., how accurately the model describes subtle visual details) is far more critical than the model's raw parameter count or general knowledge. Qwen2.5-VL's superior ability to ground fine-grained attributes into text provides the most effective supervision for our CDR stage, justifying its selection as our default generator.

S3.1.3. Effectiveness of Discrepancy Representation Constraint

We validate the design of our task-learning objective, the Discrepancy Representation Constraint (DRC), by comparing it against two standard alternatives in Table S.4.

Table S.2. Ablation of CDR conditioning components on CUB-200-2011. We compare our full attribute-centric approach against variants with degraded text or missing image context.

Conditioning Variant	R@1	R@2
Variation 1: Text Guidance Quality		
A1: Generic Class Prompt (*"A photo of a bird"*)	86.1%	91.5%
A2: No Text Guidance (Empty Input)	85.3%	90.8%
Variation 2: Image Guidance		
A3: No Context Image (Standard T2I Fine-tuning) $\mathbf{I}$	86.4%	91.7%
DiT-Distill (Full: Context + Attribute Text)	**87.2%**	**92.4%**

Table S.3. Ablation on the choice of MLLM for generating attribute descriptions. We compare models of varying sizes on CUB-200-2011.

MLLM Generator	Model Size	R@1	R@2
Keye-VL1.5 [59]	8B	86.7%	91.7%
GLM-V4.1 [16]	9B	86.9%	92.1%
Qwen2.5-VL (Ours) [2]	**7B**	**87.2%**	**92.4%**

Table S.4. Ablation study of different task-learning constraints on CUB-200-2011. We compare our proxy-based approach against standard classification and contrastive methods.

Method Type	Loss Function	R@1	R@2
Classification	Cross-Entropy Loss	80.5%	86.9%
Pairwise Metric	Contrastive Loss	85.0%	91.1%
Proxy-based Metric	**Our $\mathcal{L}_{\text{TASK}}$ (DRC)**	**87.2%**	**92.4%**

Table S.5. Ablation study comparing our Curriculum strategy against Single-Step and Feature Fusion baselines on CUB-200-2011.

Strategy	t_1 (Noisiest)	t_4 (Clearest)	**Feature Fusion**	t_3-t_4	t_2-t_4 **(Ours)**
R@1	84.4%	86.7%	85.9%	87.0%	**87.2%**
R@2	90.3%	92.0%	91.5%	92.3%	**92.4%**

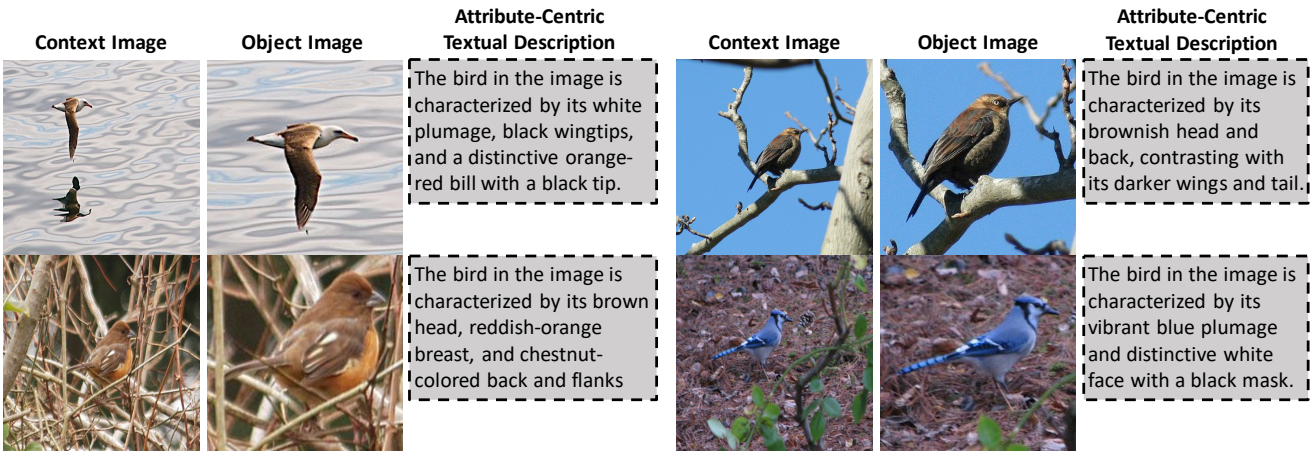

Figure S.1. Example of context-object image pairs and their corresponding attribute-centric textual descriptions on the CUB-200-2011.

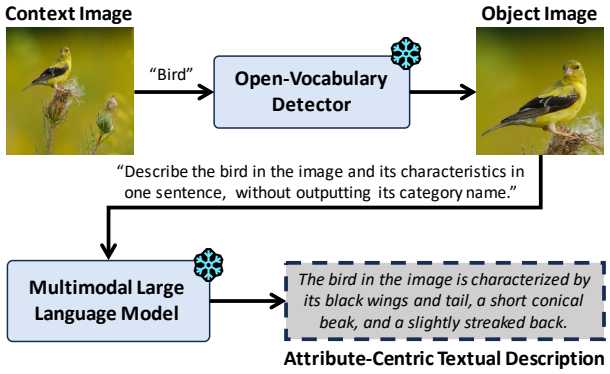

Figure S.2. An overview of the construction of context-to-object image pairs and the generation of their corresponding attribute-centric textual descriptions.

Comparison with Cross-Entropy. Standard Cross-Entropy (CE) loss yields the weakest performance (80.5% R@1). This is expected because CE optimizes for *decision boundary separability* rather than *intra-class compactness*. While it classifies training data correctly, it fails to structure the embedding space for open-set retrieval.

Comparison with Contrastive Loss. Contrastive Loss improves performance to 85.0% by explicitly optimizing pairwise distances. However, it still lags behind our method. This is likely because contrastive learning relies on local, sample-to-sample comparisons within a batch, lacking a *global* view of the class distribution.

Superiority of $\mathcal{L}_{\text{TASK}}$. Our DRC loss achieves the best performance (**87.2%**). By utilizing learnable proxies, it combines the benefits of both worlds: it provides a global context (like CE) while explicitly optimizing the cosine similarity structure (like Contrastive Loss). This forces the embeddings to form tight, discriminative clusters around stable class centers, significantly enhancing generalization to unseen subcategories.

S3.1.4. Is it Curriculum Learning or Just Feature Diversity?

A key question is whether our performance gains stem from the specific *curriculum* strategy (stage-wise alignment) or simply from accessing *diverse features* across multiple timesteps. To investigate this, we compare our method against a strong baseline:

- **Single-Step Baselines (t_1, t_4):** Distilling from a single timestep.
- **Feature Fusion Baseline:** Instead of a curriculum, we compute the average of the discrepancy embeddings from all four stages ($\bar{\mathbf{E}}_D = \frac{1}{4}\sum \mathbf{E}_{D,t_p}$) and distill this single, fused representation. This baseline has access to the *same information* as our method but lacks the *hierarchical structure*.

As shown in Table S.5, the **Fusion** baseline achieves 85.9% R@1. While this outperforms the noisy single-step t_1 (84.4%), it significantly underperforms our curriculum strategies (t_3-t_4: 87.0%, t_2-t_4: 87.2%). This -1.3% gap is revealing. If performance were driven solely by "seeing more features," Fusion should match our method. Its failure suggests that simply averaging features *dilutes* the distinct semantic information present at different granularity levels. In contrast, our curriculum approach forces the student to explicitly align with the structural progression from coarse to fine, validating that the *curriculum mechanism itself* is essential for robust representation learning.

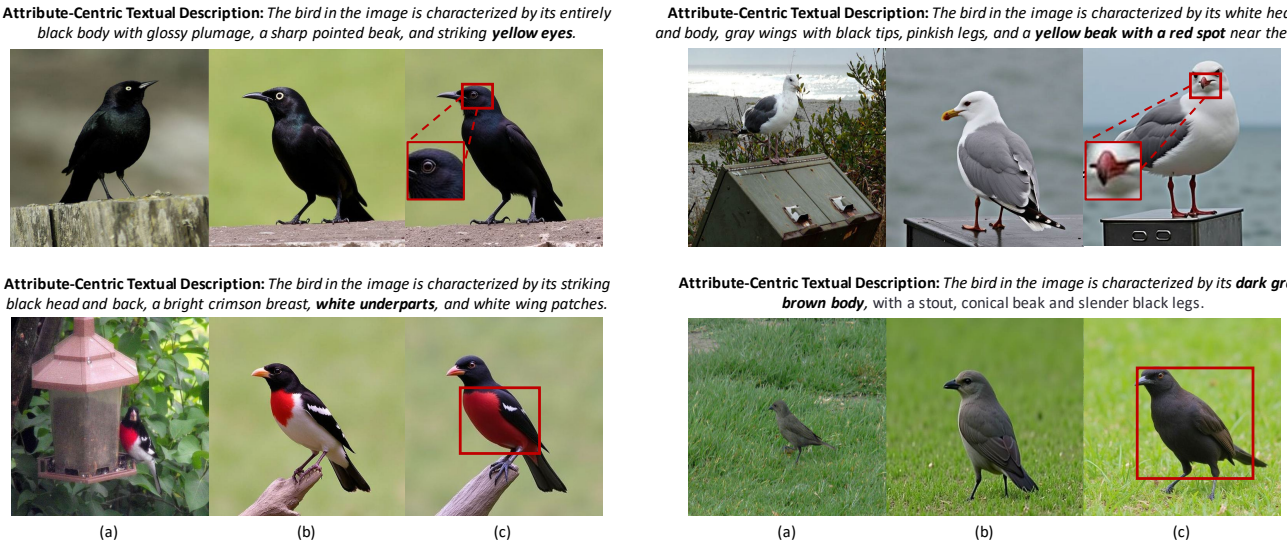

Figure S.3. Illustration of generated images. (a) shows the input (context) image, while (b) and (c) show the images generated by CDR-DiT and CDR-DiT (Mismatched), respectively.

Table S.6. Ablation study of teacher knowledge sources on CUB-200-2011. We compare our Generative Teacher (DiT) against state-of-the-art Discriminative Teachers (CLIP, DINOv2) to validate the unique benefit of generative curriculum knowledge.

Teacher Type	Teacher Model	Recall@1	Recall@2
Discriminative	DINOv2 (ViT-B/16) [40]	84.2%	90.2%
	CLIP (ViT-B/16) [43]	84.9%	90.7%
Generative	**DiT-Distill (DiT)**	**87.2%**	**92.4%**

S3.1.5. Superiority of Generative Curriculum Knowledge

A critical question is whether our performance gains stem from the specific nature of DiT's knowledge or simply from distilling a large-scale foundation model. To answer this, we replace the DiT teacher with two state-of-the-art **discriminative teachers**: CLIP [43] (text-aligned) and DINOv2 [40] (visual-centric), keeping the rest of the distillation pipeline unchanged. As shown in Table S.6, while CLIP achieves a respectable 84.9% R@1 due to its rich open-set semantics, it still lags significantly behind DiT-Distill (87.2%). This +2.3% gap highlights a fundamental difference in knowledge nature:

- **Discriminative models** (CLIP/DINO) are trained for *invariance*—ignoring subtle details to align images with texts or augmentations.
- **Generative models** (DiT) are trained for *reconstruction*—learning to synthesize every fine-grained detail from noise.

Our results quantitatively prove that for fine-grained retrieval, the constructive, detail-oriented knowledge embedded in DiT's generative curriculum is superior to the static, invariant features of discriminative giants. This validates that GCK is not just "more knowledge", but the *right kind* of knowledge for capturing subtle visual discrepancies.

Table S.7. Scalability analysis using different retrieval backbones on CUB-200-2011. Performance gains (↑) are relative to the ViT-B/16 baseline.

Backbone	Recall@1	Recall@2	Params (M)
ViT-B/16	87.2%	92.4%	86
ViT-L/14	88.7% (↑ 1.5%)	93.2% (↑ 0.8%)	307
ViT-H/14	**89.8%** (↑ 2.6%)	**93.7%** (↑ 1.3%)	632

Table S.8. Comparison with state-of-the-art Parameter-Efficient Fine-Tuning (PEFT) methods on CUB-200-2011. Our method outperforms traditional adaptation techniques by injecting generative knowledge.

Method	Base Model	Recall@1	Recall@2
CLIP-Adapter [14]	CLIP-ViT-B/16	69.8%	80.5%
AdaptFormer [8]	ViT-B/16 (IN-21k)	73.3%	82.9%
CLIP-LoRA [62]	CLIP-ViT-B/16	76.5%	84.7%
ViT-LoRA [18]	ViT-B/16 (IN-21k)	83.0%	89.2%
DiT-Distill (Ours)	**ViT-B/16 (IN-21k)**	**87.2%**	**92.4%**

S3.1.6. Scalability with Stronger Backbones

To assess the scalability of our framework, we replace the standard ViT-B/16 student with larger architectures, ViT-L/14 and ViT-H/14, as shown in Table S.7. Performance consistently improves as model capacity increases,

with ViT-H/14 reaching an impressive 89.8% R@1. Crucially, this trend confirms that our **Generative Curriculum Knowledge (GCK)** is rich and complex enough to provide additional supervision signal even for high-capacity models. It proves that DiT-Distill is model-agnostic and can scale up to leverage stronger backbones for even higher performance.

S3.1.7. Comparison with Different Parameter-Efficient Fine-Tuning Methods.

We compare DiT-Distill against standard Parameter-Efficient Fine-Tuning (PEFT) methods in Table S.8. Standard PEFT methods (e.g., ViT-LoRA, 83.0%) are limited because they only *adapt* the pre-existing knowledge within the frozen backbone. They struggle to learn new, fine-grained discrepancies that were never encoded in the original pre-training. In contrast, DiT-Distill (87.2%) fundamentally differs by *injecting* external, generative knowledge from the DiT teacher. This +4.2% lead over the strongest PEFT baseline (ViT-LoRA) demonstrates that for open-set fine-grained tasks, simply adapting discriminative features is insufficient; transferring rich, constructive details from a generative source is the key to breaking the performance ceiling.

S3.2. Visualization of CDR-DiT

This supplementary is for Sec. 4.4 of the main paper.

Validating the Context Image: What Object to Look. We first assess whether CDR-DiT can correctly identify the object in the context image and generate an object-centered image. As shown in Fig. S.3, CDR-DiT takes a context image as input and generates object-focused images (Fig. S.3 (b)) that highlight subtle visual discrepancies. These results visually confirm that CDR-DiT successfully identifies and preserves the object in the input image.

Validating the Attribute-Centric Textual Description: What Inside the Object. We validate the attribute-centric textual descriptions for CDR-DiT by examining whether they effectively guide the model to generate images consistent with the descriptions. To evaluate this, we compare CDR-DiT with a variant using mismatched textual descriptions, CDR-DiT (Mismatched). As shown in Fig. S.3, without accurate attribute-centric descriptions, the generated visual attributes are incorrect (*e.g.,* yellow eyes or a yellow beak with a red dot in the first row, Fig. S.3 (c)). These visual results confirm that attribute-centric textual descriptions convey the object's characteristics to CDR-DiT.

S4. Limitations

Despite the strong performance of DiT-Distill, we acknowledge two primary limitations.

First, the **training overhead** is significant. The CDR refinement stage requires fine-tuning a 12B-parameter DiT, which demands substantial GPU memory and time. However, it is worth noting that this is a one-time training cost; our final deployed model remains lightweight and efficient (DiT-free).

Second, our method currently relies on the **high-fidelity generation capabilities** of state-of-the-art models like FLUX [31] to source its Generative Curriculum Knowledge. The effectiveness of distilling from smaller or less capable generative models remains to be fully explored.

In future work, we aim to address these efficiency bottlenecks by investigating distillation from lightweight diffusion models and extending our "Generative Curriculum" paradigm to other fine-grained tasks such as detection and segmentation.

No. 62425603) and Basic Research Program of Jiangsu Province (Grant No. BK20240011).

References

[1] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. Cuda: Curriculum of data augmentation for long-tailed recognition. *arXiv preprint arXiv:2302.05499*, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 4, 1

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48, 2009. 1

[4] Steve Branson, Grant Van Horn, Serge J. Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014. 6

[5] Meiqi Cao, Rui Yan, Xiangbo Shu, Jiachao Zhang, Jinpeng Wang, and Guo-Sen Xie. Mup: Multi-granularity unified perception for panoramic activity recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7666–7675, 2023. 2

[6] Meiqi Cao, Xiangbo Shu, Xin Jiang, Rui Yan, Yazhou Yao, and Jinhui Tang. Exploiting frequency dynamics for enhanced multimodal event-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5969–5979, 2025. 2

[7] Huanran Chen, Yinpeng Dong, Zhengyi Wang, Xiao Yang, Chengqi Duan, Hang Su, and Jun Zhu. Robust classification via a single diffusion model. In *International Conference on Machine Learning*, 2024. 2, 3

[8] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 4

[9] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 3

[10] E Dataset. Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR*. Citeseer, 2011. 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *The International Conference on Learning Representations*, 2021. 2, 6

[12] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan V. Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7399–7409, 2022. 1, 2, 6

[13] Junyao Gao, Yanan SUN, Fei Shen, Xin Jiang, Zhening Xing, Kai Chen, and Cairong Zhao. Faceshot: Bring any character into life. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[14] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 4

[15] Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, and Tim GJ Rudner. Pre-trained text-to-image diffusion models are versatile representation learners for control. *Advances in Neural Information Processing Systems*, pages 74182–74210, 2024. 3

[16] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 2

[17] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge J. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015. 6

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *The International Conference on Learning Representations*, page 3, 2022. 2, 4, 5, 6

[19] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. DDP: diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21684–21695, 2023. 2, 3

[20] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhen-Zhong Lan, Shiguang Shan, and Alexander G. Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems*, pages 2078–2086, 2014. 1

[21] Xin Jiang, Hao Tang, Junyao Gao, Xiaoyu Du, Shengfeng He, and Zechao Li. Delving into multimodal prompting for fine-grained visual classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2570–2578, 2024. 1

[22] Xin Jiang, Hao Tang, and Zechao Li. Global meets local: Dual activation hashing network for large-scale fine-grained image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6266–6279, 2024. 2

[23] Xin Jiang, Hao Tang, Rui Yan, Jinhui Tang, and Zechao Li. DVF: advancing robust and accurate fine-grained image retrieval with retrieval guidelines. In *ACM Multimedia*, pages 2379–2388, 2024. 1, 2, 6

[24] Xin Jiang, Meiqi Cao, Hao Tang, Fei Shen, and Zechao Li. Fine-grained image retrieval via dual-vision adaptation. *CoRR*, abs/2506.16273, 2025. 2, 6

[25] Xin Jiang, Jingwen Chen, Yehao Li, Yingwei Pan, Kezhou Chen, Zechao Li, Ting Yao, and Tao Mei. Dreamvar:

Taming reinforced visual autoregressive model for high-fidelity subject-driven image generation. *arXiv preprint arXiv:2601.22507*, 2026. 3

[26] Xin Jiang, Ziye Fang, Fei Shen, Junyao Gao, and Zechao Li. Progressive feature encoding with background perturbation learning for ultra-fine-grained visual categorization. *IEEE Transactions on Image Processing*, 35:585–598, 2026. 2

[27] Xin Jiang, Hao Tang, Yonghua Pan, and Zechao Li. Rethinking vision transformer for large-scale fine-grained image retrieval. *IEEE Transactions on Multimedia*, 28:671–683, 2026. 2

[28] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3244, 2020. 6

[29] Suyeon Kim, Dongha Lee, SeongKu Kang, Sukang Chae, Sanghwan Jang, and Hwanjo Yu. Learning discriminative dynamics with label corruption for noisy label detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22477–22487, 2024. 1

[30] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[31] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3, 4, 5, 6

[32] Xiaofan Li, Yanpeng Sun, Chenming Wu, Fan Duan, YuAn Wang, Weihao Bo, Yumeng Zhang, and Dingkang Liang. Video4edit: Viewing image editing as a degenerate temporal process. *arXiv preprint arXiv:2511.18131*, 2025.

[33] Xiaofan Li, Chenming Wu, Yanpeng Sun, Jiaming Zhou, Delin Qu, Yansong Qu, Weihao Bo, Haibao Yu, and Dingkang Liang. Fvar: Visual autoregressive modeling via next focus prediction. *arXiv preprint arXiv:2511.18838*, 2025. 3

[34] Yijun Liang, Shweta Bhardwaj, and Tianyi Zhou. Diffusion curriculum: Synthetic-to-real data curriculum via image-guided diffusion. In *ICCV*, 2025. 1

[35] Jongin Lim, Sangdoo Yun, Seulki Park, and Jin Young Choi. Hypergraph-induced semantic tuplet loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 212–222, 2022. 2, 6

[36] Lizhao Liu, Shangxin Huang, Zhuangwei Zhuang, Ran Yang, Mingkui Tan, and Yaowei Wang. DAS: densely-anchored sampling for deep metric learning. In *Proceedings of the European Conference on Computer Vision*, pages 399–417, 2022. 6

[37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55, 2024. 4, 1

[38] Hongyuan Lu and Wai Lam. PCC: paraphrasing with bottom-k sampling and cyclic learning for curriculum data augmentation. In *European Chapter of the Association for Computational Linguistics*, pages 68–82, 2023. 1

[39] Yair Movshovitz-Attias, Alexander Toshev, Thomas K. Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 360–368, 2017. 1, 2, 6

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4

[41] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28285–28295, 2024. 3

[42] Xiaofan Que and Qi Yu. Dual-level curriculum meta-learning for noisy few-shot learning tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14740–14748, 2024. 1

[43] 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 *International Conference on Machine Learning*, pages 8748–8763, 2021. 4

[44] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 2, 3

[46] Jenny Denise Seidenschwarz, Ismail Elezi, and Laura Leal-Taixé. Learning intra-batch connections for deep metric learning. In *Proceedings of the International Conference on Machine Learning*, pages 9410–9421, 2021. 6

[47] Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. Cleandift: Diffusion features without noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 117–127, 2025. 3

[48] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130:108792, 2022. 2

[49] Hao Tang, Zechao Li, Dong Zhang, Shengfeng He, and Jinhui Tang. Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1958–1974, 2024. 3

[50] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, pages 1363–1389, 2023. 3

[51] Eu Wern Teh, Terrance DeVries, and Graham W. Taylor. Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis. In *The European Conference on Computer Vision*, pages 448–464, 2020. 1, 2, 6

[52] Chengkun Wang, Wenzhao Zheng, Junlong Li, Jie Zhou, and Jiwen Lu. Deep factorized metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2023. 6

[53] Haoyu Wang, Yuhu Cheng, Wei Zhang, Xiaomin Liu, and Xuesong Wang. Giddm: Generating labels with diffusion model to promote cross-domain open-set image recognition. *IEEE Transactions on Image Processing*, pages 1–1, 2025. 2

[54] Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Open-set fine-grained retrieval via prompting vision-language evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19381–19391, 2023. 1

[55] Shijie Wang, Jianlong Chang, Zhihui Wang, Haojie Li, Wanli Ouyang, and Qi Tian. Fine-grained retrieval prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2644–2652, 2023. 1, 2, 6

[56] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017. 1

[57] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Han-Jia Ye, and Jian Yang. A^2-net: Learning attribute-aware hash codes for large-scale fine-grained image retrieval. *Advances in Neural Information Processing Systems*, 34:5720–5730, 2021. 4

[58] Xiu-Shen Wei, Yang Shen, Xuhao Sun, Peng Wang, and Yuxin Peng. Attribute-aware deep hashing with self-consistency for large-scale fine-grained image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13904–13920, 2023. 4

[59] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025. 2

[60] Seonghyeon Ye, Jiseon Kim, and Alice Oh. Efficient contrastive learning via novel data augmentation and curriculum learning. In *Empirical Methods in Natural Language Processing*, pages 1832–1838, 2021. 1

[61] Zhenghang Yuan, Lichao Mou, Qi Wang, and Xiao Xiang Zhu. From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. *IEEE transactions on geoscience and remote sensing*, 60:1–11, 2022. 1

[62] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 4

[63] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2

[64] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3

[65] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Baochang Zhang, Yongjian Wu, and Feiyue Huang. Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9291–9298, 2019. 6

[66] Tianyi Zhou and Jeff Bilmes. Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity. In *International Conference on Learning Representations*, 2018. 1

[67] Tianyi Zhou, Shengjie Wang, and Jeff A. Bilmes. Curriculum learning by dynamic instance hardness. In *Advances in Neural Information Processing Systems*, 2020. 1

[68] Xiaoyu Zhu, Hao Zhou, Pengfei Xing, Long Zhao, Hao Xu, Junwei Liang, Alexander Hauptmann, Ting Liu, and Andrew C. Gallagher. Open-vocabulary 3d semantic segmentation with text-to-image diffusion models. In *The European Conference on Computer Vision*, pages 357–375, 2024. 2, 3