PanGu-Draw: Advancing Resource-Efficient Text-to-Image Synthesis with Time-Decoupled Training and Reusable Coop-Diffusion

Guansong Lu[1], Yuanfan Guo[1], Jianhua Han[1], Minzhe Niu[1], Yihan Zeng[1], Songcen Xu[1], Zeyi Huang[2], Zhao Zhong[2], Wei Zhang[1], and Hang Xu[1]

[1] Huawei Noah's Ark Lab [2] Huawei

1 More Details about PanGu-Draw

Prompt Enhancement LLM with RLAIF Algorithm. To further enhance our generation quality, we harness the advanced comprehension abilities of large language models (LLM) [5, 14] to align users' succinct inputs with the detailed inputs required by the model. Specifically, shown in Figure 3, we first construct a human-annotated dataset that enriches succinct prompts with background and

Fig. 1: Text-to-image generation results without and with prompt enhancement. Enriched text improve image generation by better image aesthetic perception (left), more detailed background (middle) and better interpretation of abstract concepts (right).

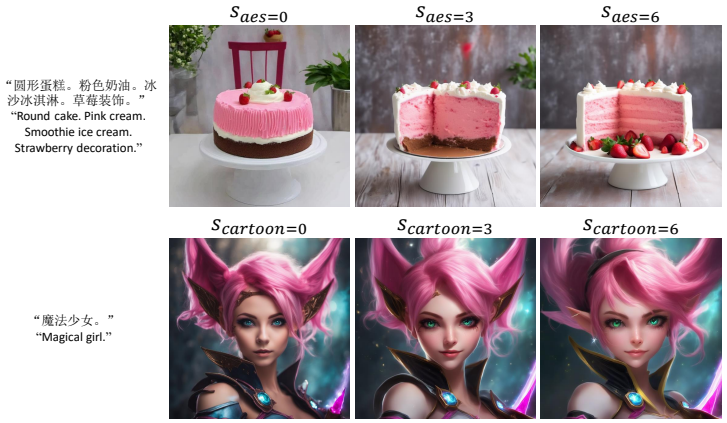


Fig. 2: Controllable stylized text-to-image generation results of PanGu-Draw. PanGu-Draw can control the generated images towards the desired style with the style guidance scale. s_{aes} for human-aesthetic-prefer style and $s_{cartoon}$ for cartoon style.

style descriptions and then fine-tune the LLM to adapt a succinct prompt to an enriched one using this data. To better adapt to the inputs required by PanGu-Draw, we perform further refinement based on the Reward rAnked FineTuning (RAFT) [4] method. Subsequently, we use the fine-tuned LLM to expand on multiple texts, which are then input into PanGu-Draw for image generation. The best expansions are selected jointly by an aesthetic scoring model[1] and a CLIP [10] semantic similarity calculation model, allowing for further fine-tuning of the LLM.

Figure 1 shows the generation results of PanGu-Draw without and with prompt enhancement. As we can see, prompt enhancement serves to add more details and illustration to the original brief prompts, leading to better image aesthetics and semantic alignment.

Controllable Stylized Text-to-Image Generation. While techniques like LoRA [7] allow one to adapt a text-to-image model to a specific style (e.g., cartoon style, human-aesthetic-preferred style), they do not allow one to adjust the degree of the desired style. To this end, inspired by the classifier-free guidance mechanism, we propose to perform controllable stylized text-to-image generation by first construct a dataset consisting of human-aesthetic-prefer, cartoon and other samples with a pretrained human aesthetic scoring model and a cartoon image classification models, and then train the text-to-image generation model with these three kinds of samples. For human-aesthetic-prefer and cartoon samples, we prepend a special prefix to the original prompt, denoted as c_{aes} and $c_{cartoon}$ respectively. During sampling, we extrapolated the prediction

[1] https://github.com/christophschuhmann/improved-aesthetic-predictor

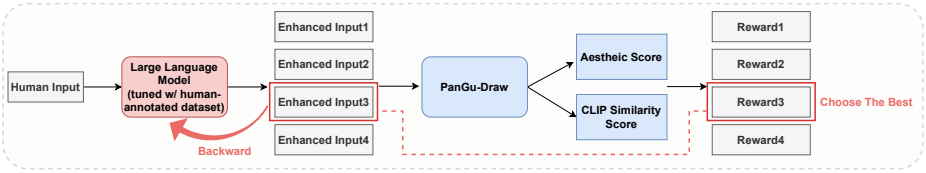


Fig. 3: Prompt enhancemnent pipeline with Large Language Model (LLM), specifically tailored for PanGu-Draw. Initially, we fine-tune the LLM using a human-annotated dataset, transforming a succinct prompt into a more enriched version. Subsequently, to optimize for PanGu-Draw, we employ the Reward rAnked FineTuning (RAFT) method, as introduced in [4], which selects the prompt pairs yielding the highest reward for further fine-tuning.

in the direction of $\epsilon_\theta(z_t, t, c_{style})$ and away from $\epsilon_\theta(z_t, t, c)$ as follows:

$$\begin{aligned}\hat{\epsilon}_\theta(z_t, t, c) = \epsilon_\theta(z_t, t, \emptyset) + s \cdot (\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \emptyset)) \\ + s_{style} \cdot (\epsilon_\theta(z_t, t, c_{style}) - \epsilon_\theta(z_t, t, c)),\end{aligned}$$

where s is the classifier-free guidance scale, $c_{style} \in \{c_{aes}, c_{cartoon}\}$ and s_{style} is the style guidance scale.

Figure 2 shows the controllable stylized text-to-image generation results of PanGu-Draw, including human-aesthetic-prefer and cartoon style image generation. As we can see, with the corresponding style guidance scale, PanGu-Draw can control the generated images towards the desired style.

2 More Implementation Details

Our models are trained on a cluster consisting of 256 Ascend 910B cards. During training, we applied several techniques to reduce redundant memory usage. These include replacing traditional attention with Flash Attention [3], employing mixed-precision training [8], and using gradient checkpointing [2], also known as the recompute technique. These methods enable the model to fit within the memory of a single Neural Processing Unit (NPU), allowing parallelism to be applied only in the data scope and avoiding model sharding among NPUs, as well as reducing inter-machine communication overhead.

3 Image Resolutions for Multi-Resolution Training

Table 1 shows the list of resolutions used for multi-resolution training of our structure generation model and texture generation model.

Table 1: The image resolutions used for multi-resolution training of structure generation model and texture generation model.

Structure Generation Model		Texture Generation Model	
Height	Width	Height	Width
512	2048	256	1024
512	1920	256	960
704	1408	384	768
768	1344	416	736
864	1152	480	640
1024	1024	512	512
1152	864	640	480
1344	768	736	416
1408	704	768	384
1920	512	960	256
2048	512	1024	256

4 More Generation Results of PanGu-Draw

4.1 Text-to-Image Generation

Figure 4 shows more generated images of PanGu-Draw. As we can see, the generated images are of high visual quality and are well aligned with the input prompts.

4.2 Multi-Diffusion Fusing Results

Multi-Control Image Generation. To benchmark the generation quality of our multi-control image generation method, we compare results shown in Figure 6 in main text with results of ControlNet with the corresponding English prompt. Results are shown in Figure 5. As we can see, both results are of similar quality. However, ControlNet does not support Chinese input text while our algorithm supports image generation conditioned on both Chinese text and pose/edge by combining PanGu-Draw with ControlNet without training a new model. As for Figure 5 in main text, the image variation model does not support text as input so one needs to train a new model conditioned on both image and text, while our algorithm makes it by combining PanGu-Draw and an image variation model without training a new model.

Figure 6 shows results of multi-control image generation by fusing PanGu-Draw with different models, including image variation, depth-to-image, edge-to-image generation models.

Figure 7 shows results of fusing two ControlNet models with our algorithm and with the algorithm proposed by ControlNet [15], which fuses the features of different ControlNets before injecting into the U-Net model. As we can see, our algorithm is able to specify the prompts of different ControlNets such that enabling a finer-grain control.

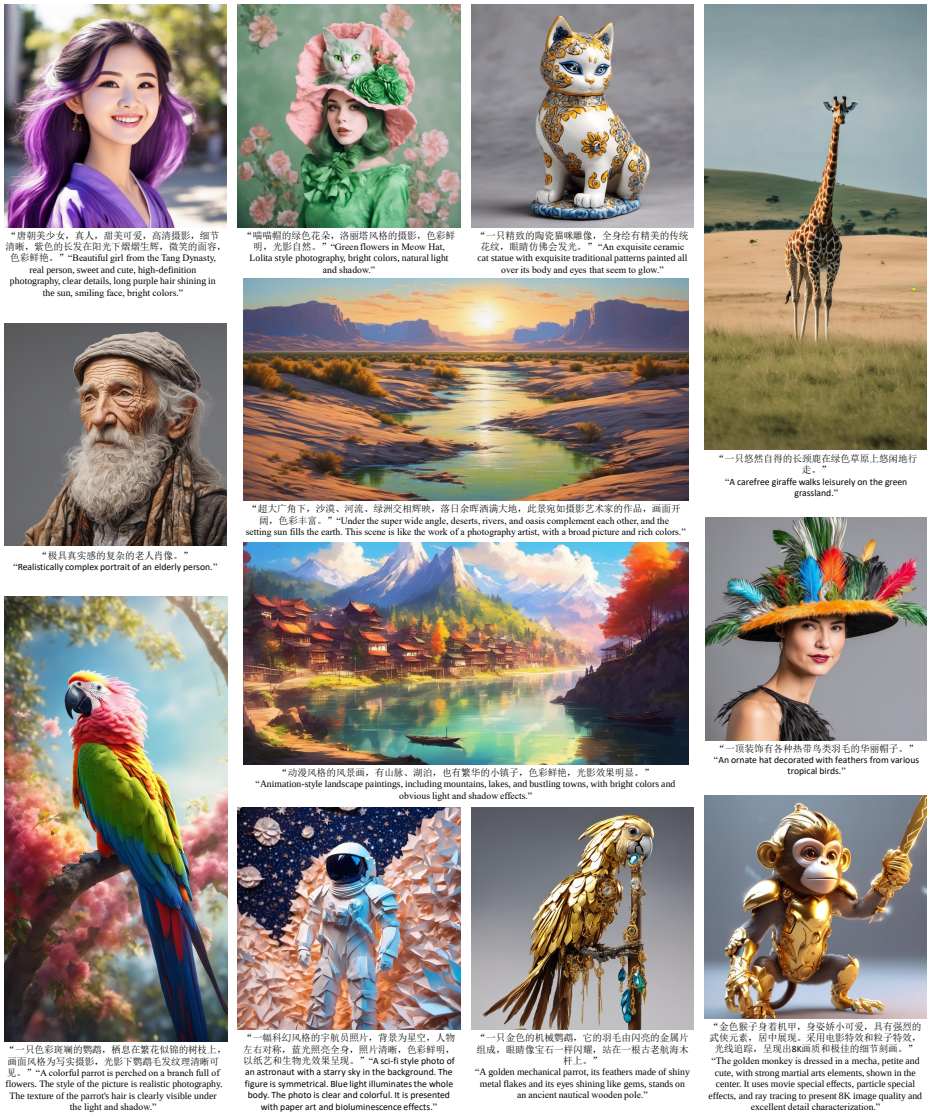


Fig. 4: Images generated with PanGu-Draw, our 5B multi-lingual text-to-image generation model. PanGu-Draw is able to generate multi-resolution high-fidelity images semantically aligned with the input prompts.

Multi-Resolution Image Generation. To benchmark the generation quality and efficiency of our *Coop-Diffusion* algorithm on single-stage super-resolution, we compare our algorithm with applying StableSR on the generated images of SD1.5. We test on 500 samples from the COCO dataset. As shown in Table 2, our algorithm achieves better quality and inference speed.

(a) English text&pose-to-image. (b) English text&edge-to-image.

Fig. 5: Generation results of Different ControlNet models.

Table 2: Comparison with SD1.5+StableSR on COCO dataset.

Method	FID↓	IS↑	CLIP-score↑	Speed (s/item)↓
SD1.5+StableSR	109.72	17.47	31.60	12.72
Coop-Diffusion(ours)	**106.80**	**18.62**	**32.10**	**10.11**

Figure 8 shows the results from the low-resolution model and our fusing algorithm *Coop-Diffusion* by fusing the low-resolution model and our high-resolution PanGu-Draw model. As we can see, PanGu-Draw adds much details to the low-resolution predictions leading to high-fidelity high-resolution results.

5 Visual Comparison against Baselines

Figure 9 and 10 shows qualitative comparisons of PanGu-Draw against baseine methods, including RAPHAEL [13], SDXL [9], DeepFloyd [12], DALL-E 2 [11], ERNIE-ViLG 2.0 [6], PixArt-α [1] and . The input prompts are also used in RAPHAEL and are provided at the bottom of the figure. As we can see, PanGu-Draw generates high-quality images, which are better than or on par with these top-performing models.

6 Potential Negative Impact and Limitations

Our method offers a more efficient way to train and apply a text-to-image generation model. We do not expect a negative social impact as the generated images are generated with textual guidance. If the training data on some sensitive categories is available, the proposed method could be misused for generating images of such sensitive categories like violence and weapons. Such misuse of image generation models poses a societal threat, and we do not condone using our work with the intent of spreading misinformation or tarnishing reputation.

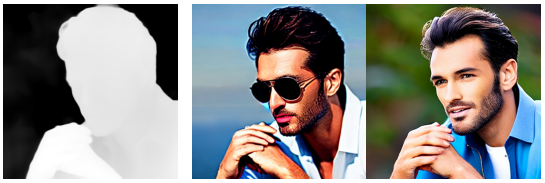
Our method currently have the following limitations: (1) As the VAE model used in our model is lossy, details of small objects like small human face might not be perfectly generated; (2) Our model is trained with English and Chinese captions, so might not work properly in other language.

(a) Input image and generation results of image variation.

"一栋黑色的房子"
("a black house")

(b) Input prompt and generation results of PanGu-Draw.

(c) Generation results of fusing image variation and PanGu-Draw with *Coop-Diffusion* algorithm.



(a) Input depth map and generation results of depth-to-image.

"一个女人"
("a woman")

(b) Input prompt and generation results of PanGu-Draw.

(c) Generation results of fusing depth-to-image and PanGu-Draw with *Coop-Diffusion* algorithm.

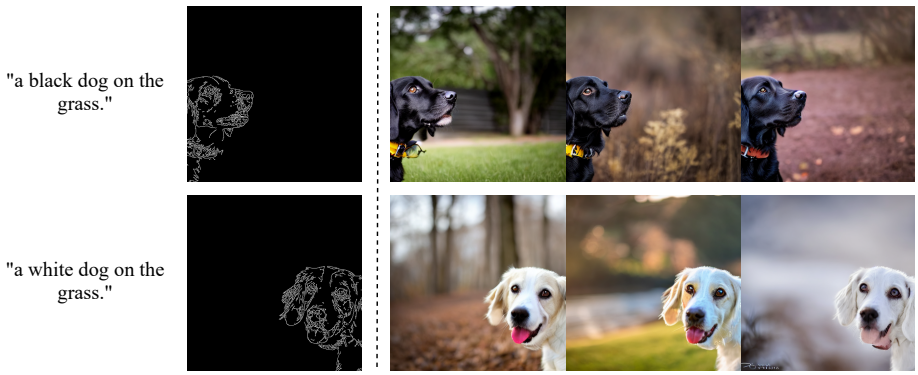
(a) Input edge map and generation results of edge-to-image.

"一栋黑色的房子"
("a black house")

(b) Input prompt and generation results of PanGu-Draw.

(c) Generation results of fusing edge-to-image and PanGu-Draw with *Coop-Diffusion* algorithm.

Fig. 6: Generation results of the fusing of an image variation/depth-to-image/edge-to-image model and PanGu-Draw with the proposed *Coop-Diffusion* algorithm.



(a) Input prompts and edge maps, and generation results of single edge-to-image ControlNet.

(b) Input prompt and generation results of Multi-ControlNets. The input edge maps are showed in (a).

(c) Generation results of fusing two ControlNets in (a) with *Coop-Diffusion*.

Fig. 7: Generation results of the fusing of an image variation/depth-to-image/edge-to-image model and PanGu-Draw with the proposed *Coop-Diffusion* algorithm.

LR model LR model + PanGu-Draw

Fig. 8: Images generated with a low-resolution (LR) model and the fusion of the LR model and HR PanGu-Draw with our *Coop-Diffusion*. This allows for single-stage super-resolution for better details and higher inference efficiency.

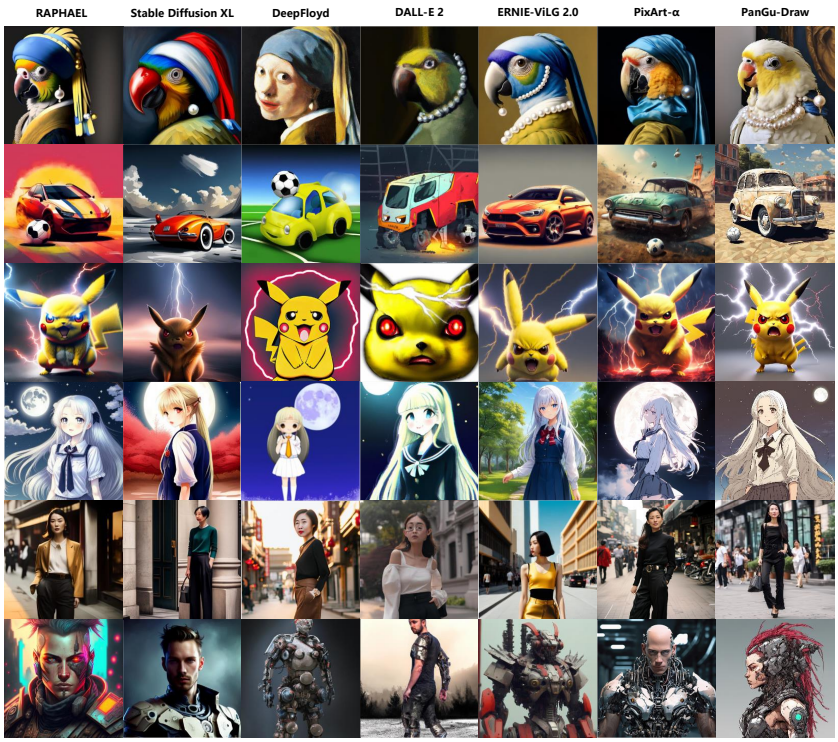


Fig. 9: Visual comparison of PanGu-Draw against baseline methods, including RAPHAEL [13], SDXL [9], DeepFloyd [12], DALL-E 2 [11], ERNIE-ViLG 2.0 [6], and PixArt-α [1]. The input prompts are also used in RAPHAEL and are provided at the bottom of the figure. The results of PanGu-Draw are better than or on par with these top-performing baseline models.

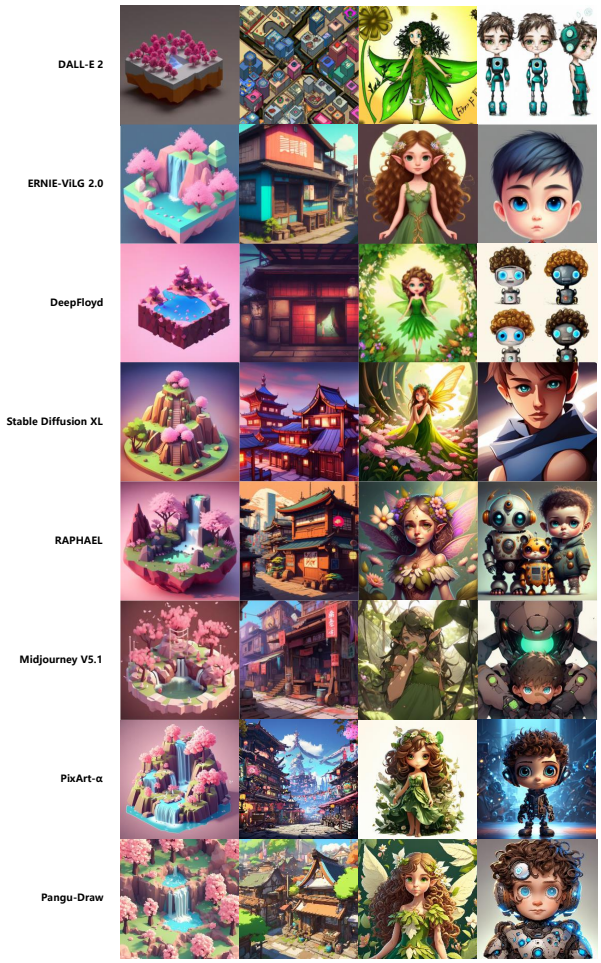


Fig. 10: Visual comparison of PanGu-Draw against baseline methods, including DALL-E 2 [11], ERNIE-ViLG 2.0 [6], DeepFloyd [12], SDXL [9], RAPHAEL [13],Midjourney V5.1 and PixArt-α [1]. The input prompts are also used in RAPHAEL and are provided at the bottom of the figure. The results of PanGu-Draw are better than or on par with these top-performing baseline models.

References

1. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
2. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost (2016)
3. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness (2022)
4. Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767 (2023)
5. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: Glm: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 320–335 (2022)
6. Feng, Z., Zhang, Z., Yu, X., Fang, Y., Li, L., Chen, X., Lu, Y., Liu, J., Yin, W., Feng, S., et al.: Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10135–10145 (2023)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
8. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training (2018)
9. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
11. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
12. Shonenkov, A., Konstantinov, M., Bakshandaeva, D., Schuhmann, C., Ivanova, K., Klokova, N.: Deepfloyd if: A powerful text-to-image model that can smartly integrate text into images (2023), `https://www.deepfloyd.ai/deepfloyd-if`, online; accessed 16-November-2023
13. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv preprint arXiv:2305.18295 (2023)
14. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)
15. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)