RomanTex: Decoupling 3D-aware Rotary Positional Embedded Multi-Attention Network for Texture Synthesis

Supplementary Material

A. Supplementary on Baselines and Experimental Criteria

Text2Tex is an inpainting-based approach, while SyncMVD and TexPainter leverage synchronizing techniques (at latent space and image space) to enhance multi-view consistency. In contrast, Paint3D employs an inpainting-dependent texture synthesis in the first stage, followed by a UV-based refinement to enhance the texture quality. On the other hand, TexGen utilizes an end-to-end UV space diffusion approach to generate textures. However, since their work relies on a incomplete UV texture as a starting point, we used a partial texture baked from the original reference image for initialization. Since most of the existing work focuses on a text-to-texture task, we improved the original SyncMVD [2] (referred to as SyncMVD-IPA) by incorporating the SDXL-base model [4] and an additional Image IP-adapter [7] to align with an image-to-texture task and compared it with our approach. Specifically, leveraging the Clean-FID [3] implementation, we harnessed a CLIP-version of Fréchet Inception Distance FID_{CLIP} to compute the distance re-renderings and the ground-truth renderings. Besides, the recently proposed CLIP Maximum-Mean Discrepancy (CMMD) [1] is also utilized to serve as an complementary criteria to validate the distribution of generated texture. In addition to these two metrics, we also use CLIP-I / CLIP-T score [5] to validate semantic alignment between renderings of the generated texture map and given image / text and LPIPS [8] to estimate the consistency between renderings of the generated texture map and the reference images.

B. Supplementary on 3D-aware RoPE

Inspired by Rotary Position Embedding (RoPE) [6] in language modeling, we inject positional information by rotating hidden state vectors according to the corresponding voxels, expressed as

$$f(x_m,(x,y,z)) = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_m \\ x_{m+1} \\ \vdots \\ x_{2m+1} \\ x_{2m+2} \\ \vdots \\ x_{3m+3} \end{pmatrix} \otimes \begin{pmatrix} \cos x\theta_0 \\ \cos x\theta_0 \\ \vdots \\ \cos x\theta_{\lfloor m/2 \rfloor} \\ \cos y\theta_{\lfloor (m+1)/2 \rfloor} \\ \vdots \\ \cos y\theta_m \\ \cos z\theta_{m+1} \\ \vdots \\ \cos z\theta_{\lfloor (3m+3)/2 \rfloor} \end{pmatrix} + \begin{pmatrix} x_1 \\ -x_0 \\ \vdots \\ -x_{m-1} \\ x_{m+2} \\ \vdots \\ -x_{2m} \\ x_{2m+3} \\ \vdots \\ x_{3m+2} \end{pmatrix} \otimes \begin{pmatrix} \sin x\theta_0 \\ \sin x\theta_0 \\ \vdots \\ \sin x\theta_{\lfloor m/2 \rfloor} \\ \sin y\theta_{\lfloor (m+1)/2 \rfloor} \\ \vdots \\ \sin y\theta_m \\ \sin z\theta_{m+1} \\ \vdots \\ \sin z\theta_{\lfloor (3m+3)/2 \rfloor} \end{pmatrix} \tag{1}$$

C. Supplementary on Ablation Study

To further evaluate the effectiveness of 3D-aware RoPE numerically, we introduced a criterion called the local alignment distance (LAD). This score computes the average Mean Squared Error (MSE) over overlapping regions between adjacent views, providing a quantitative measure of multiview coherence. The LAD is defined as follows:

$$\text{LAD} = \sum \left|\left| M_v^{UV} \odot \left\{ T_v^{UV} - \left[\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} T_v^{UV} \odot M_v^{UV} \right] \right\} \right|\right|^2 \tag{2}$$

where T_v^{UV} and M_v^{UV} represent the texture and mask in UV space unwrapped from the image of view v. As demonstrated in Suppl. C, our 3D-aware RoPE significantly outperforms the naive self-attention mechanism that lacks 3D geometry awareness.

Method	LAD
w/o MVA	0.142
w/o 3D-aware RoPE	0.123
w/o 3D-aware RoPE	**0.119**

Table 1. Ablation study on local alignment distance (LAD)

References

[1] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *CVPR*, pages 9307–9315, 2024. 1

[2] Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1

[3] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022. 1

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[5] 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. PMLR, 2021. 1

[6] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

[7] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 1

[8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1