CTR3D: Cross-view Token Reduction for Dense Multi-view Generation

Supplementary Material

6. Discussion and limitation

In this paper, we propose a method for generating dense-view images from a single input image. Our approach enables the formation of images and normals from 12 distinct views. Notably, CTR3D represents the first work capable of producing 12 views of images and normals simultaneously with nearly the same memory usage and running speed compared with existing method.

Performance justification. While full attention theoretically yields the highest quality, its quadratic complexity severely limits the view number. Our method overcomes this bottleneck, enabling significantly more views (12 vs. 6). This provides richer geometric information, leading to more accurate and complete reconstruction (e.g., Figure 7).

Multi-view generation from text-generated images. We performed multi-view generation from text-generated images and evaluated CLIP scores on renderings of the reconstructed 3D meshes. As shown in Tab. 5, our method achieved the best scores compared to baselines. Fig. 7 further illustrates our method's effectiveness, showing how generating more views (including non-horizontal) results in more complete meshes. Additionally, we also compared our method with the recent baseline method LaRa [5] which uses Zero123++ [48] to generate multi-view images from text-generated images, and then reconstructs these multi-view images into 3D Gaussians. As shown in Tab. 5 and Fig.8, our method achieves better novel-view generation than LaRa.

Compared with existing token reduction methods. Although numerous methods have been proposed to accelerate token reduction in transformers [18, 19, 23], we found that these approaches are generally unsuitable for multi-view generation. This is primarily because, in addition to accelerating reduction, it is essential to restore the number of tokens to ensure high-fidelity image generation. In our experiments, we not only compared our method with ToMe but also implemented other learnable reduction techniques [46, 69] that incorporate reduction-recuperation mechanisms. However, we observed that these methods typically incur significant overhead when restoring the number of tokens. This issue is particularly pronounced in high-resolution multi-view scenarios, where the number of tokens increases exponentially. Consequently, the training costs associated with these methods are prohibitively high, and our computational resources are insufficient to support their training. As a result, we have excluded comparisons with these methods from our experimental results.

Hyperparameter analysis. Note that our current choice of hyperparameter K (which is $[4096, 1536, 384, 96]$) in our designed model is empirical, balancing efficiency and performance, which is effective in all of our experiments. To provide a more detailed analysis, we re-evaluated the speed of our Cross-view Token Reduced Attention layer under different hyperparameter settings. Tab. 6 illustrates how different K values affect runtime across varying number of views and generation image resolutions.

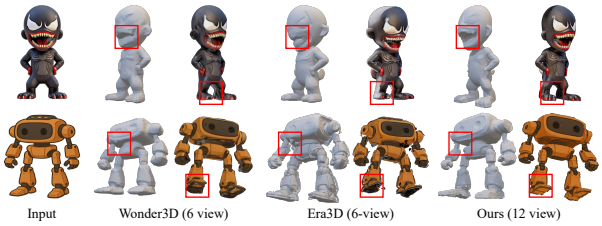

Figure 7. Qualitative comparison of text-to-image-to-multi-view.

Compared with existing native 3D generation methods. In this paper, we mainly focus on efficient high-resolution dense multi-view generation. We present a comparative analysis of novel view synthesis against baseline methods, specifically Wonder3D [39] and Era3D [25]. Additionally, we highlight the primary application of our method: reconstruction. Existing literature suggests that an increased number of views contributes to improved reconstruction outcomes, a claim supported by our findings on the GSO dataset. In this study, we focus on demonstrating the results of integrating our CTR3D with NeuS2 [56] (the same reconstruction method used in Wonder3D and Era3D) for the reconstruction task. The reconstruction comparisons underscore the significance of utilizing multiple views for effective reconstruction and illustrate the capability of CTR3D in generating multi-view consistent images and normals. Additionally, in Fig. 8, we compare the reconstruction of our generated multi-view images using NeRF versus NeuS, which indicates that NeuS presents higher capability for representing details.

It is important to note that we employed an earlier reconstruction method to facilitate quantitative comparisons with previous works, which may result in our reconstruction quality appearing inferior to that of recent native 3D generation methods [27, 59, 67]. The primary objective of this paper is to introduce a lightweight token reduction method that enhances the number of generated views based on the stable diffusion model, rather than to conduct a quality comparison of final reconstructions or generated geometries.

Furthermore, our method has the potential to com-

Figure 8. Qualitative comparison with LaRa.

Metric	LaRa	wonder3d	era3d	ours
CLIP Score ↑	0.7609	0.8206	0.8359	**0.8425**

Table 5. Quantitative comparison of text-to-image-to-multi-view.

K	(8192,3072,768,192)		(4096,1536,384,96)		(2048,768,192,48)	
(resolution, view)	(512,6)	(512,12)	(512,6)	(512,12)	(512,6)	(512,12)
Times (ms)	2.137	3.262	1.856	3.041	1.736	2.820

Table 6. Running time under different K settings.

plement recent native 3D generation techniques, such as Clay [67] and CraftsMan [27]. However, the lack of publicly available code and models for multi-view image inputs prevents us from obtaining their generated results. It is also worth noting that these methods currently support only 4 input views. While they do not yet accommodate 12-view inputs, they have demonstrated the importance and effectiveness of multi-view image inputs. With ongoing advancements, we anticipate the emergence of native 3D generation methods capable of supporting more input views in the future. At that time, the gap between generation and reconstruction is likely to narrow, producing results more aligned with the input images and further highlighting the value of our CTR3D approach.

Limitations. Due to constrained computational resources, we only conducted experiments generating 12 views. In the future, we aim to explore scenarios with denser viewpoints or higher resolutions. Additionally, our approach may occasionally produce smooth textures in the generated images. To address this, we plan to incorporate a refinement stage or add a lightweight detail-preserving branch in future work.

7. More visualized results

In this section, we present more testing results compared with Wonder3D [39] and Era3D [25]. As shown in Figure 9, our results present better completeness and quality. We argue that this improvement mainly comes from dense views and our light-weight token reduction design. The dense views provide more overlap region for reducing the complexity of reconstruction, while the light-weight token reduction promises the cross-view attention can work on dense views as well as preserving the multi-view consistency.

We also provide the generated 12 views and reconstructed mesh in Figure 10.

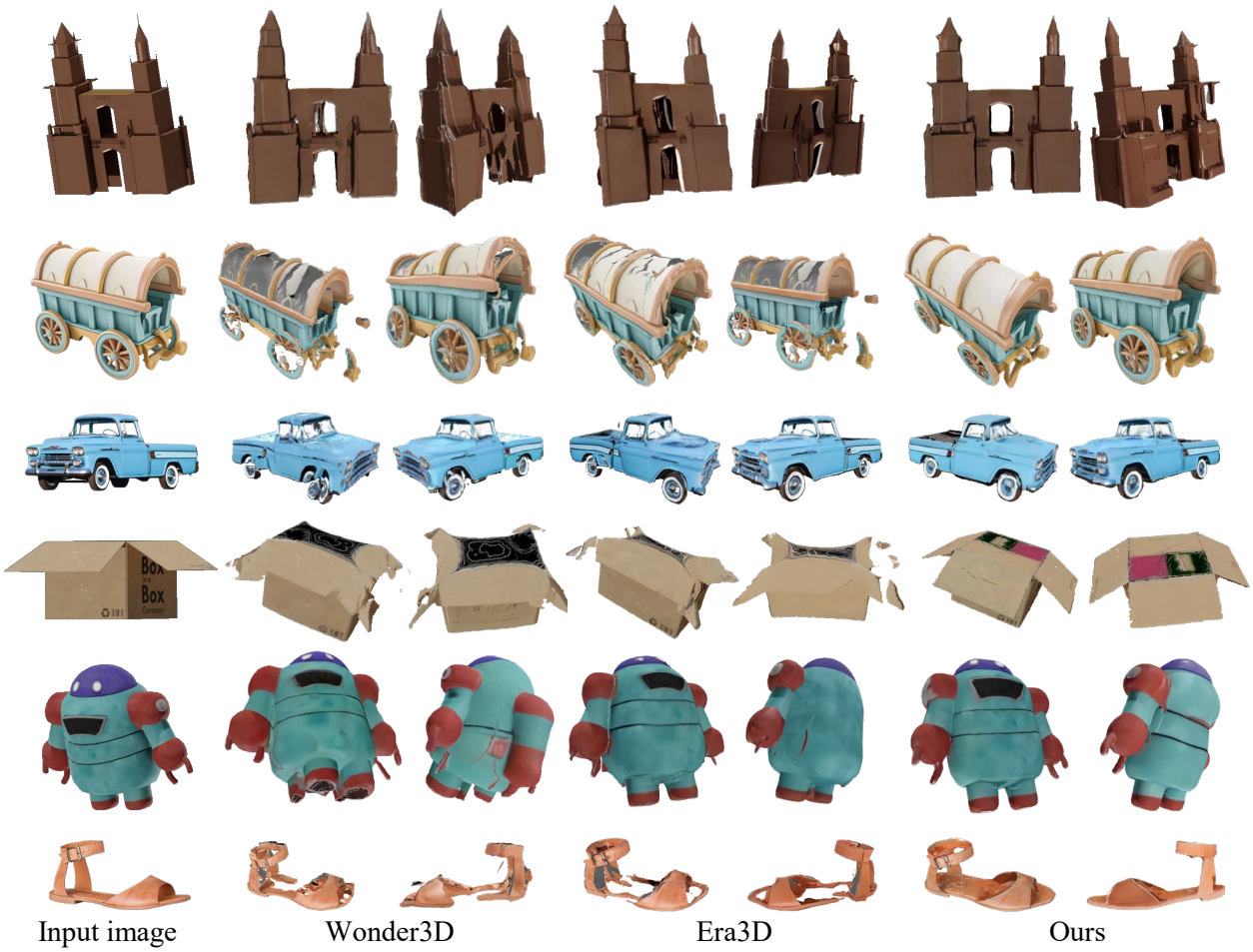

Figure 9. The visualized comparison results.

Figure 10. The visualized results generated from our CTR3D.