DDIM Inversion based Novel View Synthesis

(a) MvImgNet test set (b) Out-of-domain real-world images

Figure 1: (a) High-resolution (512×512) novel-view synthesis on the MvImgNet test set from a single input image and camera parameters, (b) Zero-shot synthesis on out-of-domain images downloaded from Unsplash.

Abstract

Synthesizing novel views from a single input image is a challenging task. It requires extrapolating the 3D structure of a scene while inferring details in occluded regions, and maintaining geometric consistency across viewpoints. Many existing methods must fine-tune large diffusion backbones using multiple views or train a diffusion model from scratch, which is extremely expensive. Additionally, they suffer from blurry reconstruction and poor generalization. This gap presents the opportunity to explore an explicit lightweight view translation framework that can directly utilize the high-fidelity generative capabilities of a pretrained diffusion model while reconstructing a scene from a novel view. Given the DDIM-inverted latent of a single input image, we employ a camera pose-conditioned translation U-Net, TUNet, to predict the inverted latent corresponding to the desired target view. However, the image sampled using the predicted latent may result in a blurry reconstruction. To this end, we propose a novel fusion strategy that exploits the inherent noise correlation structure observed in DDIM inversion. The proposed fusion strategy helps preserve the texture and fine-grained details. To synthesize the novel view, we use the fused latent as the initial condition for DDIM sampling, leveraging the generative prior of the pretrained diffusion model. Extensive experiments on MVImgNet demonstrate that our method outperforms existing methods. Code is available.

Qualitative Results

Few qualitative figures from our paper demonstrating the performance of our approach.

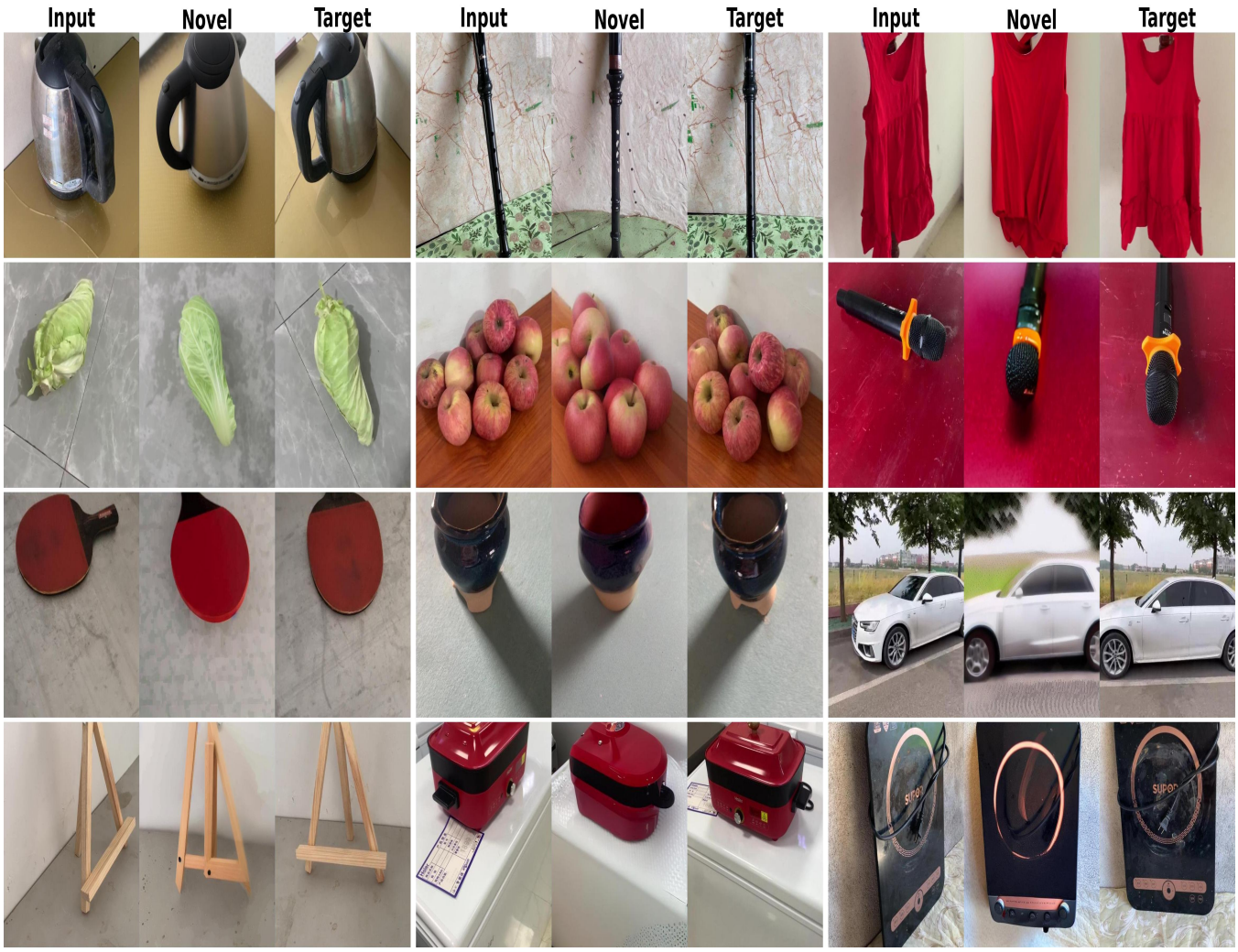

Figure 2: Qualitative results with our 167-class trained model.

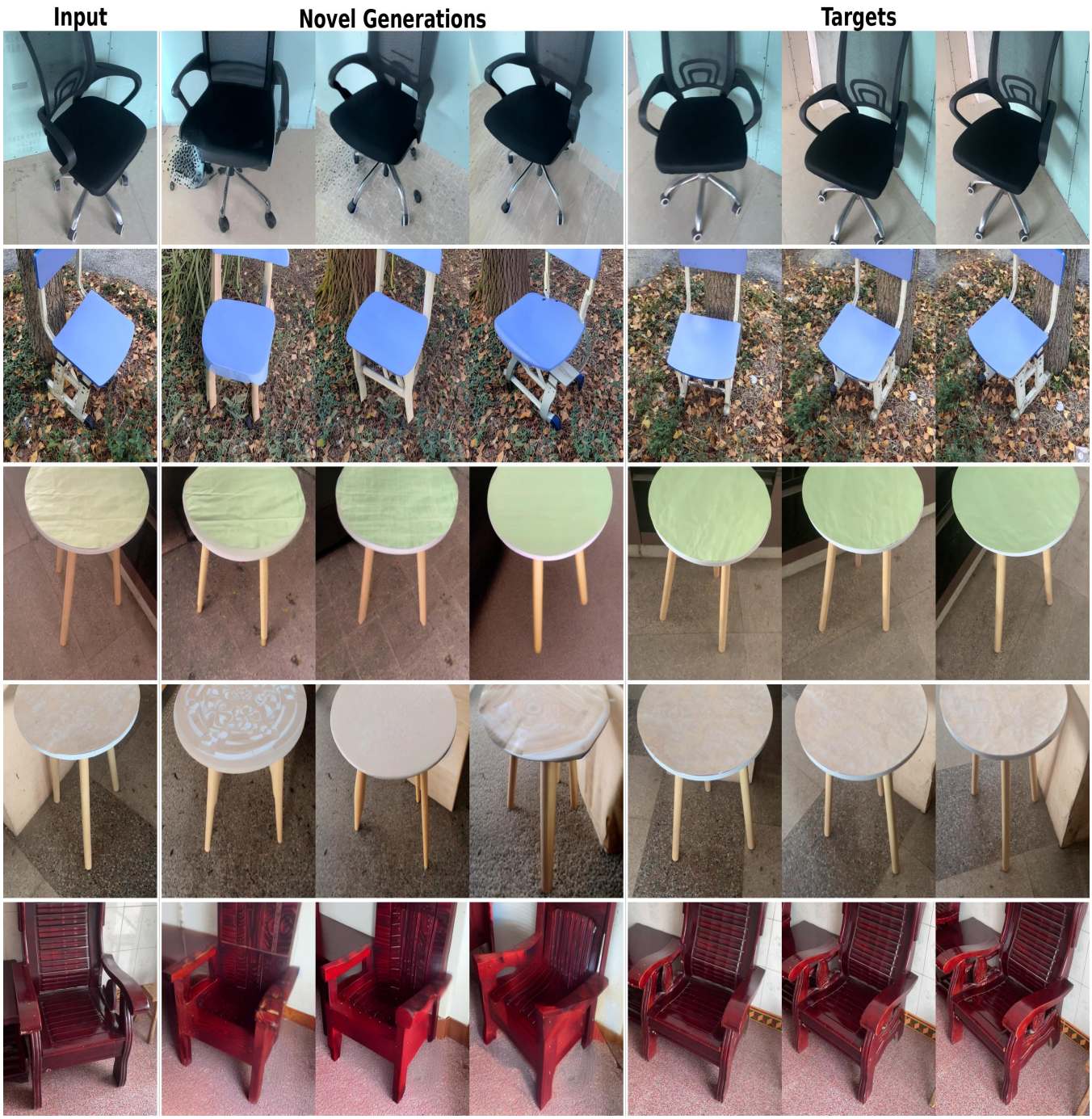

Figure 3: Generating multiple frames with single input image.