MIRAGE: Unsupervised Single Image to Novel View Generation with Cross Attention Guidance

Llukman Cerkezi[1*], Aram Davtyan[1*], Sepehr Sameni[2], Paolo Favaro[1]
[1]Computer Vision Group, University of Bern, [2]Independent Researcher, [*]Equal contribution
{llukman.cerkezi, aram.davtyan}@unibe.ch, sepehr.sameni@gmail.com, paolo.favaro@unibe.ch

1. Additional Experimental Results

In this section, we present additional results of MIRAGE that could not be included in the main paper due to space constraints. Figures 1 and 2 show further qualitative comparisons between MIRAGE and other methods. A key limitation of competitor methods is their inability to generate plausible outputs for unseen views, particularly for realistic objects such as cars, airplanes, and buses. This suggests that despite being trained on large amounts of supervised synthetic data, these methods still struggle with generalization. Figure 3 presents the 3D point clouds generated by applying Dust3R [5] to both the input image and the novel view for various objects. As observed, the resulting point clouds effectively capture the underlying geometry from the two views. This demonstrates that the novel views generated by MIRAGE are sufficiently consistent to produce valid and realistic 3D point clouds.

Pose-Centric clustering results. In Figures 4, 5, 6, 7, 8, and 9 we show visually Pose-Centric clustering results when applied for multiple category dataset. It is worth noting that we found a trade-off between the number of clusters and the pureness of the poses inside the cluster. More specifically, when we attempted to reduce the number of pose clusters, we found that images within clusters started to exhibit more diverse (noisy) poses/orientations.

3D aware GAN-based results. Lastly, we also show some samples from GIRAFFE (Fig. 10), which is also a generative model to create novel views of objects. We took the pretrained GIRAFFE model on the CompCars dataset and generated 360° views. The generated views do not exhibit a strong multi-view consistency, as both the cars' identities (shapes) and their textures change with the viewpoint. This is a well know problem for 3D-aware Generative Adversarial Networks (GANs), "as none of them can preserve strict multi-view consistency, partially on account of the usage of a 2D upsampler and lack of explicit 3D supervision" (as quoted from [2]).

2. Implementation Details

We train our models on images of size 224×224.Note that the images were padded with boundary values to ensure a uniform square shape, as they originally had varying dimensions. While 224×224 is standard resolution for many vision benchmarks, our method is resolution-agnostic and can be applied to higher-resolution inputs without architectural changes. During training, we adhere to the hyperparameters outlined in DDIM [4] [1], *i.e.*, setting the learning rate to 2×10^{-5} and utilizing the Adam optimizer [1]. We adopt the U-Net architecture from this source [2]. To make the architecture pose-conditioned, we incorporate learnable category and pose embeddings into the time embedding as we have multi-category dataset. The model is trained using eight NVIDIA GeForce RTX 3090 GPUs, and for 6.5×10^5 iterations. Note that we apply our cross-attention framework solely in the bottleneck and the decoder of the U-Net, as we found that using it in the encoder does not have an effect. We apply HAG at the inference time only starting from the 20th timestep onward (we apply 100 denoising timesteps during the inference). We exploit the 224 resolution version of Dust3r [5].

3. Discussions

We train our models from scratch because, to the best of our knowledge, existing diffusion models for novel view synthesis are trained in a supervised manner. Fine-tuning these models would therefore conflict with our objective of maintaining an unsupervised training setup.

Generating novel views in a completely unsupervised manner comes with a slight drawback that is evaluating the quality of these views using standard metrics like PSNR or SSIM becomes difficult. This is because when a novel view is generated for a specific pose ID, we do not have control over the camera parameters. Additionally, within a given pose ID cluster, there is always some variation in the ob-

[1] https://github.com/ermongroup/ddim
[2] https : / / github . com / filipbasara0 / simple - diffusion/tree/main/model

ject's orientation and scale. These factors make it very challenging to compare generated novel views with ground truth images in pixel space, even when ground truth is available.

On the other hand, the absence of camera poses requires one to manually set the parameters to obtain explicit 3D from our generated views. This issue could in theory be addressed by estimating a 3D template from average DINOv2 features at the pose cluster centroids. We leave this for future exploration. Lastly, there is room for improvement in the consistency of the generated multi-view data.

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 1

[2] Leheng Li, Qing Lian, Luozhou Wang, Ningning Ma, and Ying-Cong Chen. Lift3d: Synthesize 3d training data by lifting 2d gan to 3d generative radiance field. In *CVPR*, 2023. 1

[3] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 12

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[5] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1

Figure 1. Visual comparison of novel views generated by MIRAGE and Free3D, EpiDiff and LGM on various objects.

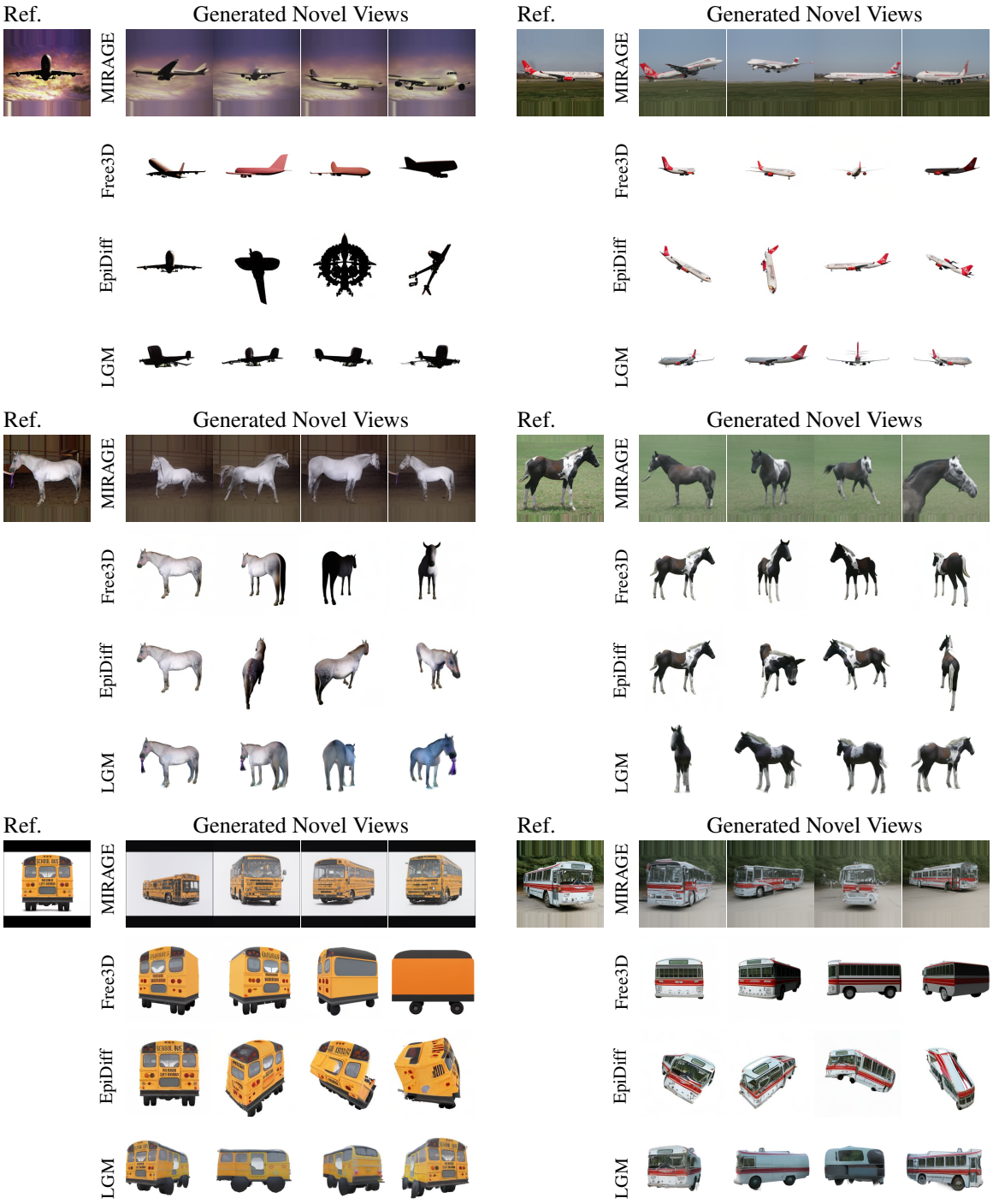

Figure 2. Visual comparison of novel views generated by MIRAGE and Free3D, EpiDiff and LGM on various objects.

Figure 3. Generated 3D point clouds by applying Dust3r to the input image and the novel view. Please zoom into the point cloud figures to see more details.

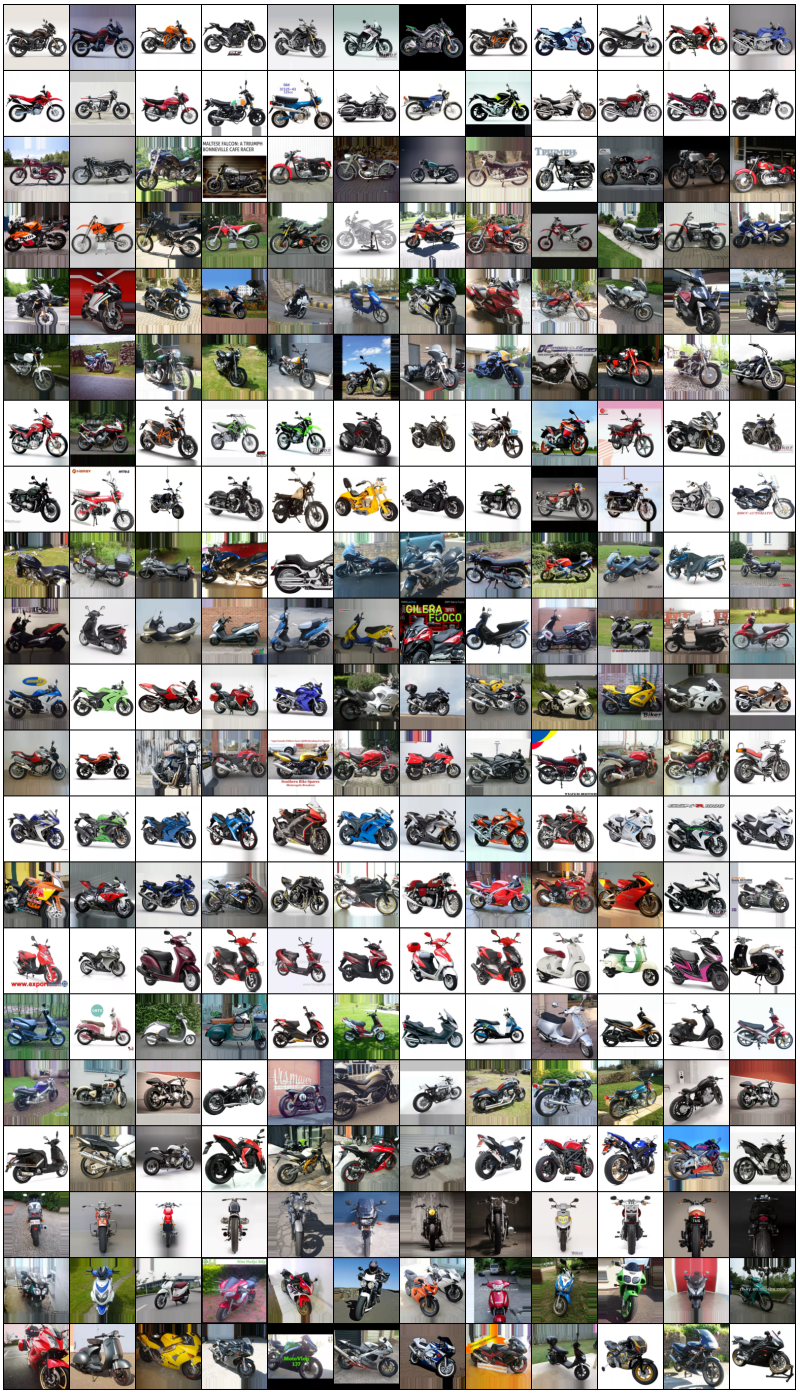

Figure 4. Each row represents images belonging to a specific pose ID.

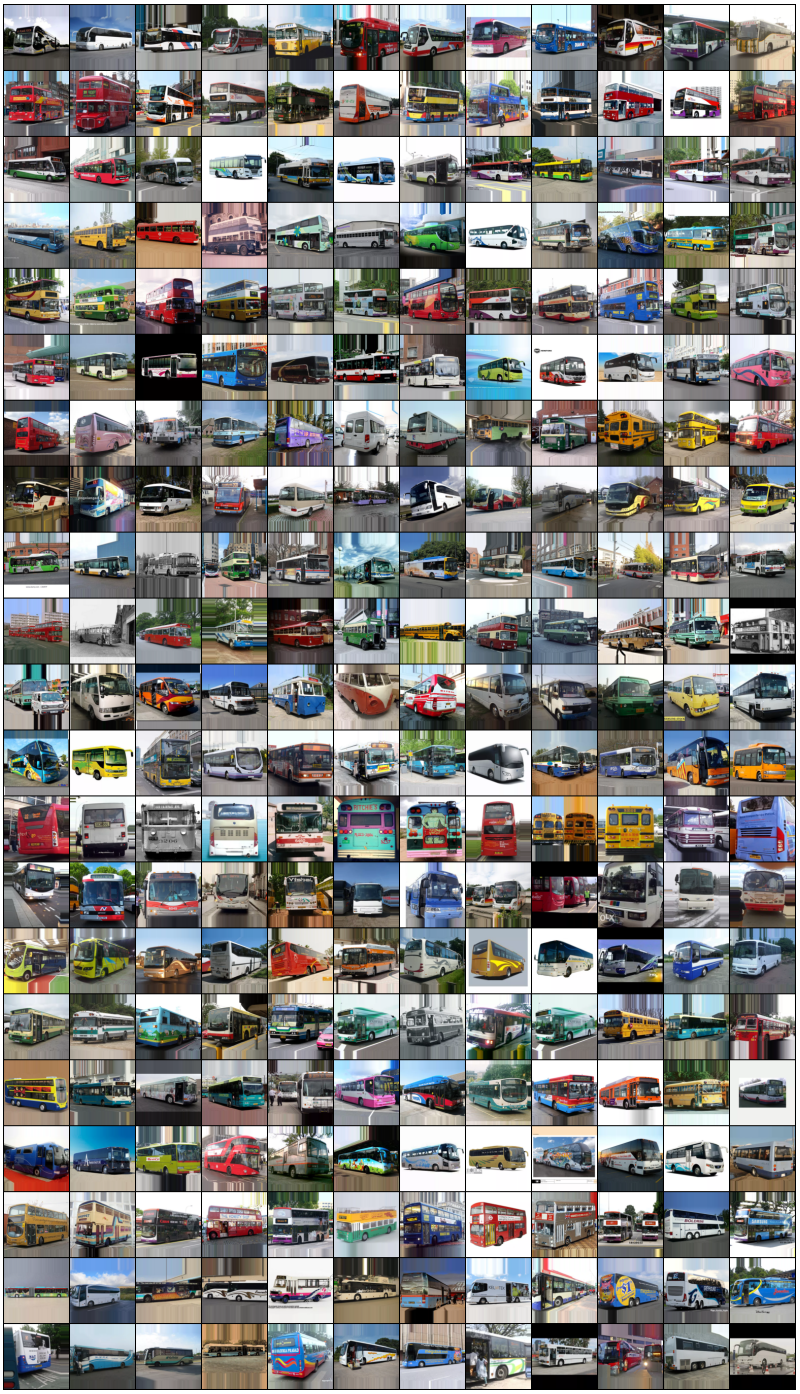

Figure 5. Each row represents images belonging to a specific pose ID.

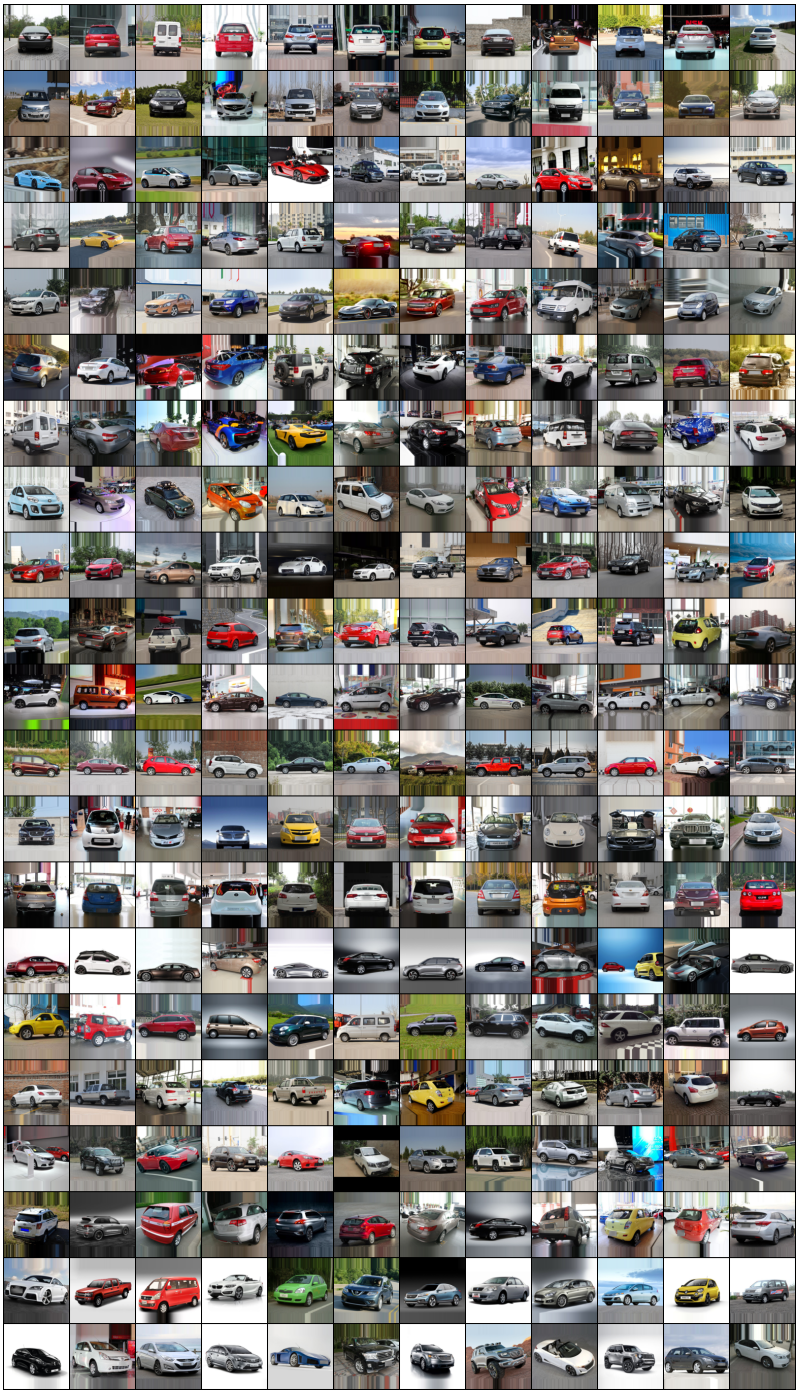

Figure 6. Each row represents images belonging to a specific pose ID.

Figure 7. Each row represents images belonging to a specific pose ID.

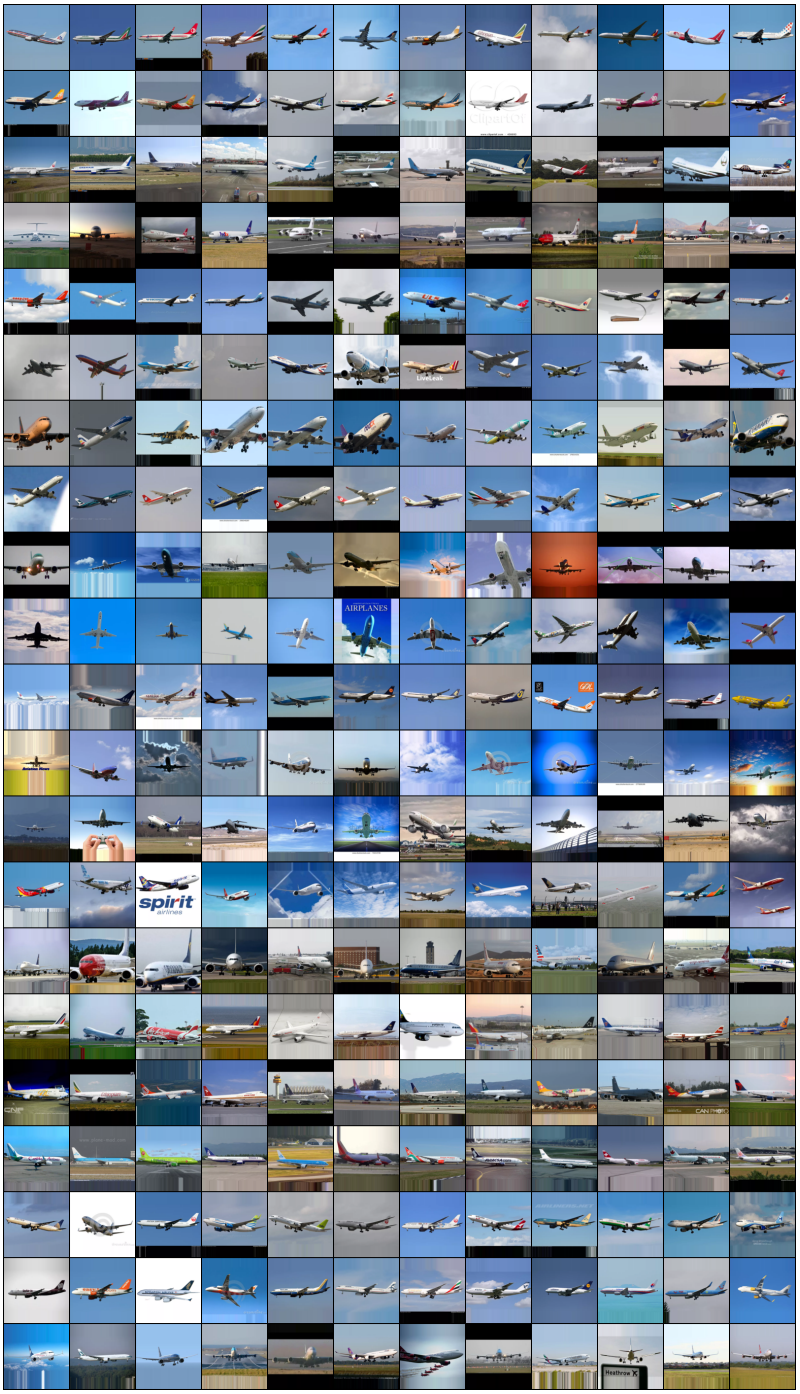

Figure 8. Each row represents images belonging to a specific pose ID.

Figure 9. Each row represents images belonging to a specific pose ID.

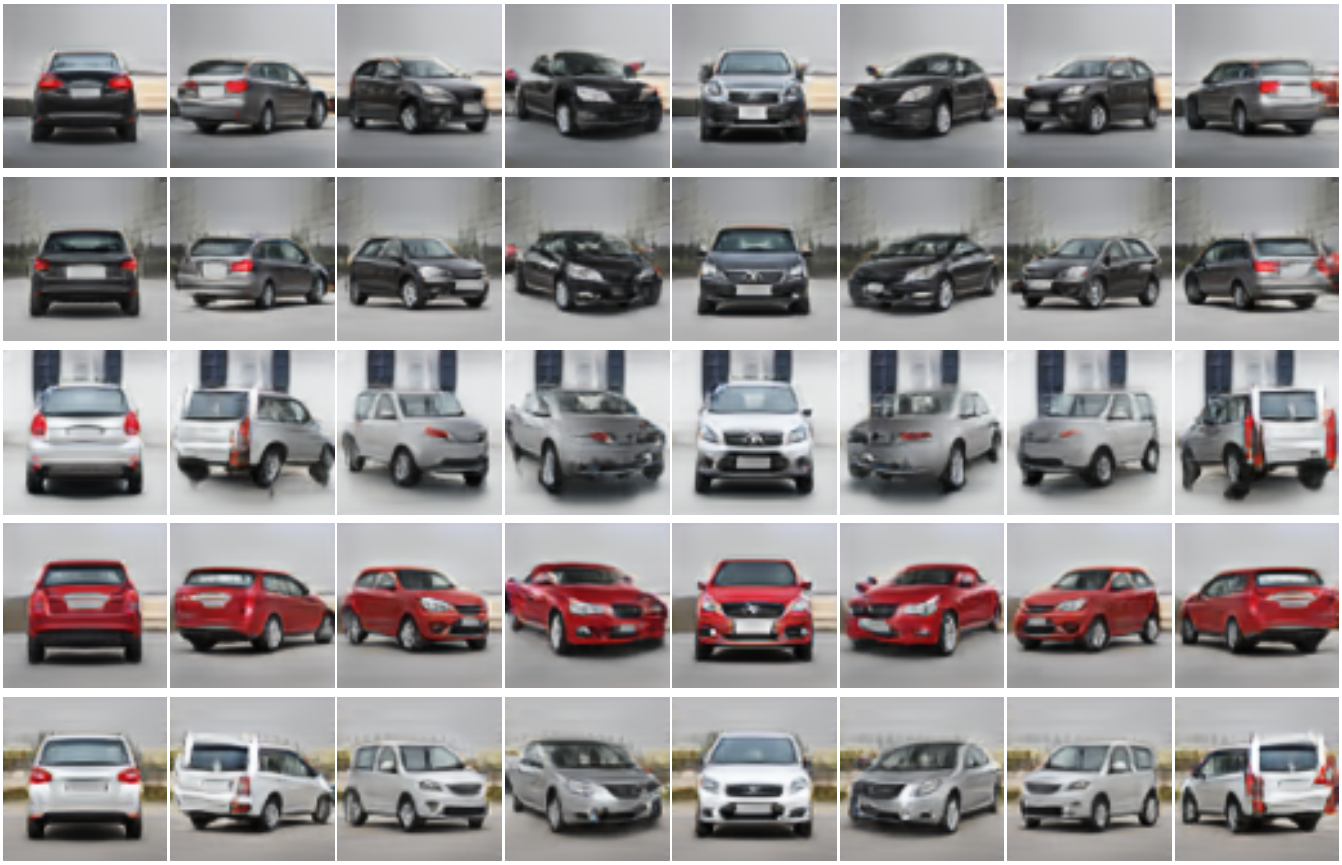

Figure 10. For comparison, 360° generated cars using GIRAFFE [3].