

# Multi-View Unsupervised Image Generation with Cross Attention Guidance

Llukman Cerkezi\* Aram Davtyan\* Sepehr Sameni Paolo Favaro  
 Computer Vision Group, Institute of Computer Science, University of Bern, Switzerland  
 {llukman.cerkezi, aram.davtyan, sepehr.sameni, paolo.favaro}@unibe.ch



Figure 1. Examples of novel views generated by MIRAGE on several datasets.

## Abstract

The growing interest in novel view synthesis, driven by Neural Radiance Field (NeRF) models, is hindered by scalability issues due to their reliance on precisely annotated multi-view images. Recent models address this by fine-tuning large text2image diffusion models on synthetic multi-view data. Despite robust zero-shot generalization, they may need post-processing and can face quality issues due to the synthetic-real domain gap. This paper introduces a novel pipeline for unsupervised training of a pose-conditioned diffusion model on single-category datasets. With the help of pretrained self-supervised Vision Transformers (DINOv2), we identify object poses by clustering the dataset through comparing visibility and locations of

specific object parts. The pose-conditioned diffusion model, trained on pose labels, and equipped with cross-frame attention at inference time ensures cross-view consistency, that is further aided by our novel hard-attention guidance. Our model, MIRAGE, surpasses prior work in novel view synthesis on real images. Furthermore, MIRAGE is robust to diverse textures and geometries, as demonstrated with our experiments on synthetic images generated with pre-trained Stable Diffusion.

## 1. Introduction

In recent years, the field of novel view synthesis has gained significant interest, primarily due to advancements in Neural Radiance Field (NeRF) models [2, 33, 35, 65]. While NeRFs have demonstrated remarkable capabilities in gen-

\*Equal contribution.

erating novel views, they require hundreds of images precisely annotated with exact camera poses. This makes these models difficult to scale efficiently in practice.

To address this challenge, an alternative approach has emerged with the development of diffusion models [19]. Specifically, a large conditional diffusion model is trained on a substantial multi-modal dataset of real images (e.g. Stable Diffusion [42]). This model is further fine-tuned on synthetic datasets containing multi-view images with associated camera poses to enable novel view synthesis [31, 47, 60]. Despite exhibiting robust zero-shot generalization, these models need additional post-processing at inference time, which involves object segmentation to remove the background. Moreover, the quality of generated novel views might still be far from satisfactory due to the domain gap between synthetic and real data. However, fine-tuning these models with real data remains unfeasible due to the impracticality and cost associated with obtaining accurate pose annotations for diverse object categories. In this work we tackle this challenge by proposing a novel pipeline based on training a pose-conditioned diffusion model on a single-category dataset in a fully unsupervised manner.

Our proposed approach starts by identifying the variety of different object poses represented in the dataset using recently developed self-supervised Vision Transformers (ViTs [11]), namely DINOv2 [37]. DINOv2 has demonstrated the ability to capture semantic relationships among object parts in images of similar or related categories. Our method utilizes this capability by comparing the visibility and the location of specific object parts in images, effectively clustering the original dataset into groups that share the same pose/orientation of objects.

Following this, we train a pose-conditioned diffusion model using the pose labels obtained in the first stage. Once trained, this diffusion model allows the generation of images featuring objects in diverse and controlled poses. However, these generated images do not yet form multi-view sequences. To impose consistency across images with different poses, we employ the cross-frame attention procedure [25]. This procedure facilitates communication of self-attention keys and values in the denoising backbone of the diffusion from the reference pose to the others. Additionally, we introduce hard-attention guidance to enhance the quality of the generated views and ensure further consistency.

We refer to our model as MIRAGE for Multi-view unsupervised Image generation with cRoss-Attention GuidancE. To evaluate our methodology, we conduct tests on a real dataset of car images. Our results demonstrate that the generated novel views not only surpass those produced in prior work but also exhibit sufficient quality for reconstructing an explicit 3D. Furthermore, to showcase the robustness of MIRAGE across various categories in the absence of diverse

real datasets, we train it on datasets generated with a pre-trained Stable Diffusion model [42].

Our contributions can be summarized as follows:

- A novel way to discover the set of poses represented in a single-category dataset;
- A pose-conditioned diffusion model that is capable of generating images of objects in predefined poses;
- An adaptation of cross-frame attention to novel view synthesis field with our novel hard-attention guidance;
- A dataset of synthetic single-category images generated with Stable Diffusion to show robustness of our method to different textures and geometries.

## 2. Literature Review

**NeRF-based Novel view synthesis.** Neural radiance fields (NeRF) [33] have proven to be a robust framework for generating novel views. Despite achieving notable success [2, 3, 7, 35, 45], these models face challenges when tasked with generating novel views from a limited number of input images. This limitation arises from the original method’s optimization, which is scene-specific. One strategy to address this limitation involves pretraining the model across multiple scenes using synthetic data to acquire a scene prior. During inference, a sparse set of views, or even a single view, is then used to generate novel views [6, 8, 15, 30, 41, 52, 63]. Note that, their inference is reliant on pretraining with multiview images and corresponding pose labels. Moreover, the quality diminishes when the target view significantly deviates from the input views, resulting in blurry generated views. For this, they primarily focus on two cases 1) generating narrow views [6, 8, 41], and 2) generating 360° views where the input image is masked and synthetic, and the input views cover almost all the texture (from the top view) [15, 30, 52, 63]. In contrast, our work tackles a more challenging task by aiming to generate 360° views for real images with background and without supervised training.

**Diffusion-based Novel View Synthesis.** In pioneering work, 3DiM [54] developed a pose conditional image-to-image diffusion model, training the network by providing the source view and relative pose to predict the target view. However, a significant limitation is its reliance on synthetic images, as the diffusion models are exclusively trained on synthetic datasets. Addressing this limitation, Zero-1-to-3 [31] proposed fine-tuning the pre-trained multi-modal diffusion model, Stable Diffusion [42], on a large-scale synthetic dataset, Objaverse [10]. While this fine-tuning allows zero-shot generalization to out-of-distribution datasets and in-the-wild images, our attempts to generate novel views for real cars, specifically from a *side* view, revealed challenges in hallucinating novel parts for other views and achieving multiview consistency. Several works, including Consistent-1-to-3 [61], Consistent123 [55], and

MVDREAM [48], aimed to enhance multiview consistency in Zero-1-to-3. These works share a common methodology of modifying self-attention layers in the U-Net [43] during training, initially proposed for generating temporally consistent videos [20, 25, 56] and consistent image synthesis and editing [4]. Specific modifications include epipolar guided attention [61], shared self-attention [55], and 3D attention [48]. In contrast, our approach differs in two key aspects. Firstly, we adjust the self-attention layer only during inference, not during training. Secondly, our novel pipeline does not require multi-view images with corresponding poses, demonstrating the ability to train with approximate and noisy discrete pose labels.

**Single-view 3D Reconstruction.** In contrast to novel view synthesis, single-view 3D reconstruction methods primarily focus on predicting the 3D geometry and appearance beyond the visible, rather than generating novel views. Common representations for 3D include mesh-based approaches [14, 22, 34, 49], volumetric methods [9, 13, 16, 62], and neural implicit 3D representations [29, 39]. A shared characteristic among these methods is their reliance on mask supervision to predict the object’s shape, with the exception of Monnier *et al.* [34]. Monnier *et al.* achieve this by enforcing consistency between images of different object instances, ensuring similarity in shape or texture and utilizing multi-stage progressive training. Although they generate reasonably good 3D meshes, the texture tends to be coarse, and the training process is cumbersome due to multi-stage progressive training. In our approach, we aim to produce more detailed and high-quality renderings of novel views while maintaining consistency. Furthermore, our method is easier and simpler to train, eliminating the need for any supervision.

**3D-aware Generative Models.** With the development of NeRF [33], many 3D-aware generative adversarial networks having radiance fields as a representation are proposed [5, 36, 46, 50, 58, 66]. Although they generate high-quality outputs, they often lack multiview consistency (shape and texture change when the camera moves) and are sensitive to prior pose estimation. This is partially due to 2D upsampler that eases the training and lack of explicit 3D supervision. In contrast, diffusion-based 3D-aware generative models have demonstrated superiority over GAN-based counterparts [23, 24], especially in generating 360° views. However, a significant drawback is their dependence on multiview datasets with known poses, such as Co3Dv2 [40], for training. In our work, we also show that we can obtain 3D-aware diffusion models without requiring multiview dataset. It is noteworthy that diffusion models offer an advantage over GAN-based models in terms of inversion, making it easier to generate query input images back, thus enabling them to be also utilized for novel views.

### 3. Method

Let  $\mathcal{D}$  be a collection of RGB-images capturing objects of the same category in different orientations. That is,  $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$ , where  $x^{(i)} \in \mathbb{R}^{3 \times H \times W}$ . In this work we assume that each  $x^{(i)}$  is an image of a single salient object. Furthermore, under general assumptions,  $\mathcal{D}$  is not necessarily a multi-view dataset, i.e. it does not have to contain multiples views of the same instance. The task of novel view synthesis from a single reference image is to generate a set of images of the object in the reference frame in different orientations/poses that are consistent with the reference.

Since synthesising novel views is ambiguous due to occlusions, we propose a probabilistic approach to this task. Let  $\mathcal{P}$  be a finite set of poses/orientations that the objects in  $\mathcal{D}$  can have, with  $|\mathcal{P}| = m$ . Given a reference image  $x_{\text{ref}}$  the task is to generate a set of images  $x_{p_1}, \dots, x_{p_m}$ , such that  $x_{p_j}$  captures the object in  $x_{\text{ref}}$  in the pose  $p_j \in \mathcal{P}$ . To this end, we model the following conditional distribution

$$p(x_{p_1}, \dots, x_{p_m} | x_{\text{ref}}) = \prod_{j=1}^m p(x_{p_j} | x_{\text{ref}}). \quad (1)$$

Thus, each  $x_{p_j}$  is sampled independently from the other views. For the separate single view conditional distributions in 1 we train a pose-conditioned diffusion [19] and resolve the conditioning on  $x_{\text{ref}}$  through sharing keys and values in the self-attention layers [53] of the denoising U-Net [43] backbone. Recently it has been shown that this technique ensures consistency across generated images in such applications as video generation [25] and style transfer [1]. Indeed, although the independent sampling of the generated views may introduce inconsistencies, in the experiments section, we demonstrate empirically that the novel views obtained with our method are sufficiently good for 3D reconstruction.

#### 3.1. Pose-Centric Clustering via Feature Visibility

So far we assumed that the set of possible poses  $\mathcal{P}$  is given. However, in the absence of such annotations one needs to discover  $\mathcal{P}$  from  $\mathcal{D}$  in an unsupervised way. To do so, we start by noticing that the pose of an object is defined by the set of visible semantic features and their locations in the image. For instance, when a car rotates, different parts of it (lights, wheels, doors, etc.) become visible at different locations, while other parts remain occluded. At the same time, for two different cars in the same pose after centering and rescaling certain parts appear in roughly similar locations. Thus, with respect to some metric that is based on comparing semantic local features of the objects after centering and rescaling the images,  $\mathcal{D}$  splits into clusters that approximately correspond to different poses in the dataset.

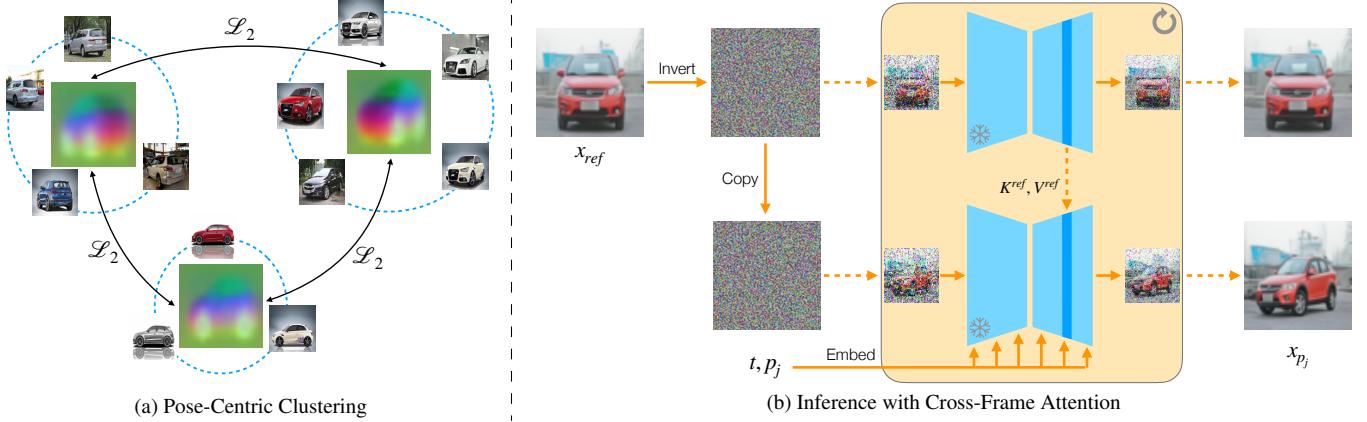


Figure 2. The illustration of the proposed pipeline for unsupervised novel view synthesis. (a) Pose-centric clustering based on  $\mathcal{L}_2$  distance between first 3 principal components of DINOv2 spatial features after centering and rescaling. The average feature map is shown in the center of each cluster, upsampled for illustration purposes. (b) The inference with MIRAGE. The keys and values in the self-attention layers of the denoising backbone of the pose-conditioned diffusion are passed from the reference pose to generate the image of the object in the reference frame in pose  $p_j$ .

As the feature extractor we adopt DINOv2 [37], as its features were shown to capture semantics of object parts. As suggested in the original paper, we start by fitting PCA [12] in the space of DINOv2 spatial features. Selecting the tokens that project to the positive axis of the first principal component allows us to roughly segment the object of interest in the image. Using the segmentation masks we center and rescale the images so that the objects span a shared predefined bounding box. We continue by fitting another PCA, this time only on the masked features (i.e. features that belong to the objects). We then project the original features to the first 3 principal components of the second PCA. This way we minimize the information about the texture in the features, while retaining the semantics of object parts (see Figure 2a). The resulting tensors are compared in euclidean distance to cluster the dataset via K-Means clustering [32]. Figure 2a showcases sample clusters after applying the described procedure. Since running PCA on the whole dataset is time-consuming, we opt for Incremental PCA [44] that works at the batch-level and refines the approximate global PCA at each iteration.

### 3.2. Pose-Conditioned Diffusion

Given the pose-cluster assignments obtained with the procedure described in 3.1, we train a diffusion that generates images conditioned on those pose labels. More precisely, we train DDPM [19] that learns to predict the noise added to the clean image by minimizing the following objective:

$$\mathcal{L}_{DDPM}(\theta) = \sum_{t=1}^T \mathbb{E}_{x_0, \varepsilon} \|\varepsilon_\theta^{(t)}(x_t; p) - \varepsilon\|^2, \quad (2)$$

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \varepsilon, \quad (3)$$

where the expectation with respect to  $x_0$  is taken over the dataset  $\mathcal{D}$ ,  $\varepsilon$  is sampled from the standard normal distribution, and  $p \in \mathcal{P}$  is the pose cluster assignment of  $x_0$ .  $\{\alpha_t\}_{t=1}^T$  is a sequence of decreasing positive real numbers defined such that the distribution of  $x_t$  is approximately standard normal as  $t \rightarrow T$  with a sufficiently large  $T$ .

At inference, we leverage deterministic DDIM sampler [51] for efficiency. DDIM starts by sampling  $x_T$  from  $\mathcal{N}(0, 1)$  and iteratively denoises it by directly predicting  $x_0$  from  $x_t$  at each iteration and using it to predict  $x_{t-1}$  via the following formula:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left( \frac{x_t - \sqrt{1 - \alpha_t} \varepsilon_\theta^{(t)}(x_t; p_j)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \varepsilon_\theta^{(t)}(x_t; p_j). \quad (4)$$

A typical choice for  $\varepsilon_\theta^{(t)}$  is the U-Net architecture [43] with group normalization [57] and built-in self-attention layers [53]. We condition  $\varepsilon_\theta^{(t)}$  on the desired pose label  $p_j$  by first embedding it to some vector space and then using this embedding to normalize the features in the residual blocks of the U-Net, similar to the AdaIN framework [21].

### 3.3. Multi-View Generation

While the conditioning on  $p_j$  enables the diffusion introduced in Section 3.2 to generate images of objects in diverse

and controlled poses, for novel view synthesis one needs to make sure that the generated images capture the same object. This is not guaranteed even when the images are generated starting from the same noise (see top row of Figure 3). Therefore, we propose to utilize an inference scheme similar to those used for video generation to ensure zero-shot temporal consistency of pretrained text-to-image diffusion models [25].

**Cross-attention is all you need.** It has been shown recently that the self-attention layers in the U-Net backbone of pretrained diffusion models capture structure (queries and keys) and appearance (values) of the objects in the image. Moreover, transferring keys and values from one image to another results in appearance transfer, while the global structure is preserved [1]. This is achieved with the so called cross-frame attention procedure. Normally the self-attention layers take feature maps  $f \in \mathbb{R}^{c \times h \times w}$  from the previous blocks of the network, linearly project them and reshape to obtain  $Q, K$  and  $V \in \mathbb{R}^{n \times d}$ , where  $n = h \times w$ , and compute the output using the following formula:

$$\text{Self-Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V. \quad (5)$$

In the cross-frame attention the self-attention layers are modified to allow attending to the keys from the reference image only. More precisely, let  $Q^{p_j}$  denote the queries calculated in one of the self-attention layers during the generation process with the pose label  $p_j$  and  $K^{\text{ref}}, V^{\text{ref}}$  be the corresponding key and values for the reference image. Then the output of the cross-frame attention (CFA) is calculated as:

$$\text{CFA}(Q^{p_j}, K^{\text{ref}}, V^{\text{ref}}) = \text{Softmax}\left(\frac{Q^{p_j}(K^{\text{ref}})^\top}{\sqrt{d}}\right)V^{\text{ref}}. \quad (6)$$

By using this technique we ensure that the global structure (i.e. the pose) of the object is aligned with the condition  $p_j$ , while the appearance, the local stucture and the identity are carried over from the reference image. This allows to generate valid novel views.

**Hard-attention guidance.** Moreover, we incorporate classifier-free guidance [18] into the denoising process. At each iteration, we replace  $\varepsilon_\theta^{(t)}(x_t; p_j)$  in equation 4 with  $\varepsilon^g$ , calculated as follows:

$$\varepsilon_t^g = (1 - \gamma) \cdot \varepsilon_t^{\text{hard}} + \gamma \cdot \varepsilon_t^{\text{soft}}, \quad (7)$$

Here,  $\varepsilon_t^{\text{soft}}$  denotes the output of the denoising network using cross-frame attention, and  $\varepsilon_t^{\text{hard}}$  follows the same procedure as  $\varepsilon_t^{\text{soft}}$  but with Argmax instead of Softmax in equation 6. We call this procedure Hard-Attention Guidance,

Dataset	Method	FID $\downarrow$
CompCars [59]	Ours - no MV	6.081
	Ours - MV, w/ bcg, w/o HAG	9.320
	Ours - MV, w/ bcg, w/ HAG	9.079
	Monnier <i>et al.</i> [34] - w/o bcg	115.560
SD Armchairs	Ours - MV, w/o bcg	34.796
	Ours - no MV	5.744
	Ours - MV, w/o HAG	7.870
	Ours - MV, w/ HAG	11.204

Table 1. The FID scores for CompCars and SD Armchairs for various versions of MIRAGE (ours) and Monnier *et al.* [34] method. MV refers to multi-view and bcg is an abbreviation for background.

or HAG. It is worth noting that, in contrast to conventional classifier-free guidance, equation 7, with  $\gamma \geq 0$ , defines a negative guidance. In our case  $\gamma = 1$  means no HAG is applied. This negative guidance directs the generation process away from situations where similar parts of the object hardly attend to different parts in the reference image, resulting in inconsistent artifacts in the generated sequences. This technique not only improves the quality of generated images but also increases consistency across views, as demonstrated in our experiments (see Figure 3).

**Single image to multi-view.** To generate new views of a given image, the initial step involves inverting the diffusion process to extract the noise responsible for generating that particular sample. For this, we exploit the deterministic nature of the DDIM [51] sampler, which allows a straightforward inversion. Once the latent noise is identified, it is used to generate other views with poses that are different from the one in the reference image. Throughout this novel view generation process, keys and values are always transferred from the reference image. This approach guarantees that the generated views maintain coherence with the input image.

## 4. Experiments

In this section, we present implementation details and results. For more comprehensive ablation studies as well as for more visual results, we refer the reader to the supplementary material.

### 4.1. Miscellaneous

**Datasets.** To assess the performance of our method we carry out experiments on the **CompCars** [59] dataset that contains up to 100K images of cars captured from various poses ranging in 360 degrees. Lacking large and diverse datasets of single-category images, in order to assess the robustness of our method to textures and shapes other than cars, we propose to leverage Stable Diffusion [42] (SD) to generate pseudo-real datasets of that nature. More precisely,

we condition SD on prompts of the kind “*A high-quality photo of (category), (view-point)*”, where the view-point can be “*front view*”, “*side view*” and “*back view*”. Additionally, we specify negatives prompts as “*Low-quality, close-up, cropped, [view-point]*”, where the optional viewpoint field is set to “*front view*” for the positive prompts featuring back and side views. In this paper we consider the “*single massive armchair*” category, while other categories are to be explored in the future work. We generate 100k images of armchairs that form the **SD Armchairs** dataset. We will release this dataset upon publication.

**Implementation Details.** For both datasets, we train our models on images of size  $64 \times 64$ . During training, we adhere to the hyperparameters outlined in DDIM [51], i.e., setting the learning rate to  $2 \times 10^{-4}$  and utilizing the Adam optimizer [26]. During inference, we set the guidance strength to  $\gamma = 1.5$  for all datasets and employ 50 denoising steps. 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. In the case of the CompCars dataset, we opt for a cluster count of 25, while for armchairs, we choose 10 clusters. In particular, we observed that in both cases, some clusters shared the same pose. 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 clusters, we found that images within clusters started to exhibit more diverse (noisy) poses/orientations.

**Baseline.** To the best of our knowledge, our method is the first approach that proposes novel view synthesis without the reliance on any form of supervision. In our evaluation, we compare with the most closely related method, specifically that of Monnier et al. [34], which reconstructs objects in an unsupervised manner. For the comparisons on the car dataset, we utilize their official pretrained models\*. However, in the case of the armchair dataset, we encountered difficulties in training their model. We hypothesize that hyperparameters play a crucial role in the convergence of their model and experiments with various random seeds are required to achieve optimal results.

## 4.2. Ablations

In this section we ablate some of our design choices. We start by emphasizing the importance of using CFA and HAG for obtaining consistent and smooth novel views (see Figure 3). Increasing the strength of HAG smoothens the results. However, too large values of  $\gamma$  result in unrealistic images. Besides this, we also noticed that the global statistics of the images are encoded in the initial noise that is used to generate the views. Therefore, starting the generation from the same noise improves consistency.

\*<https://github.com/monniert/unicorn>

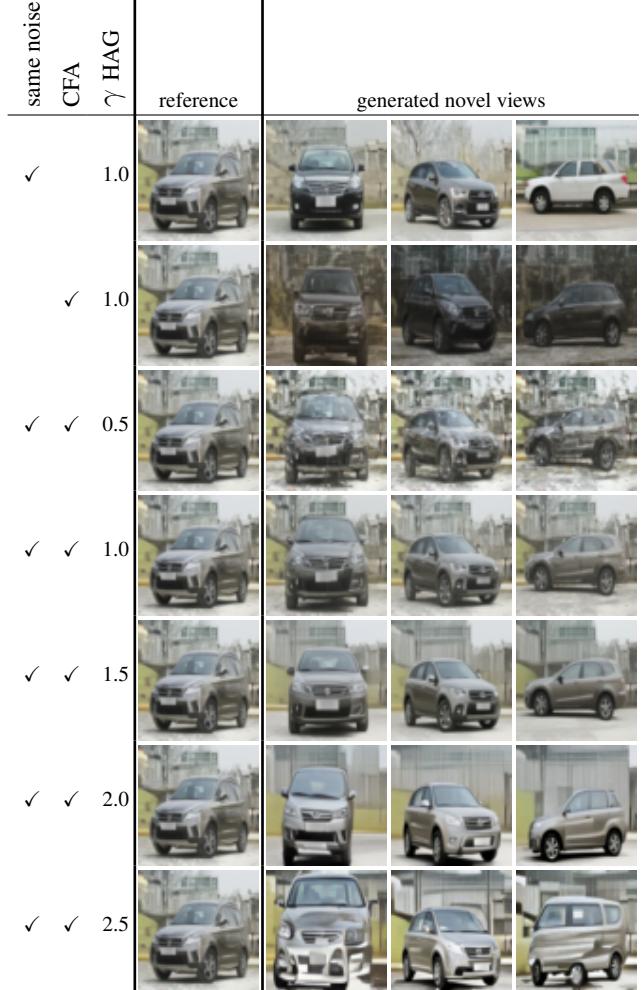


Figure 3. Ablations of the usage of the same initial noise at generation, cross-frame attention (CFA) and the strength of hard-attention guidance ( $\gamma$  HAG). The combination of CFA and generation from the same noise ensures consistency across view, while the usage of HAG imposes smoothness in the generated views and removes inconsistent artifacts.

## 4.3. Results

We present the FID scores [17] in Table 2 for both datasets. The FID no-MV indicates that the images are generated without cross-attention, i.e., in a standard way, and they have the lowest FID score in both datasets among other scores. We also computed the FID for our generated multiview dataset. More specifically for each reference image (noise) we generated 3 novel views. For fairness, we randomly selected the pose reference for each instance (noise). We generated the multiview dataset for both w and w/o guidance. For the CompCars dataset we observe the version w/ guidance having lower FID, while for the SD Armchairs dataset the results are the opposite. We hypothesize this to

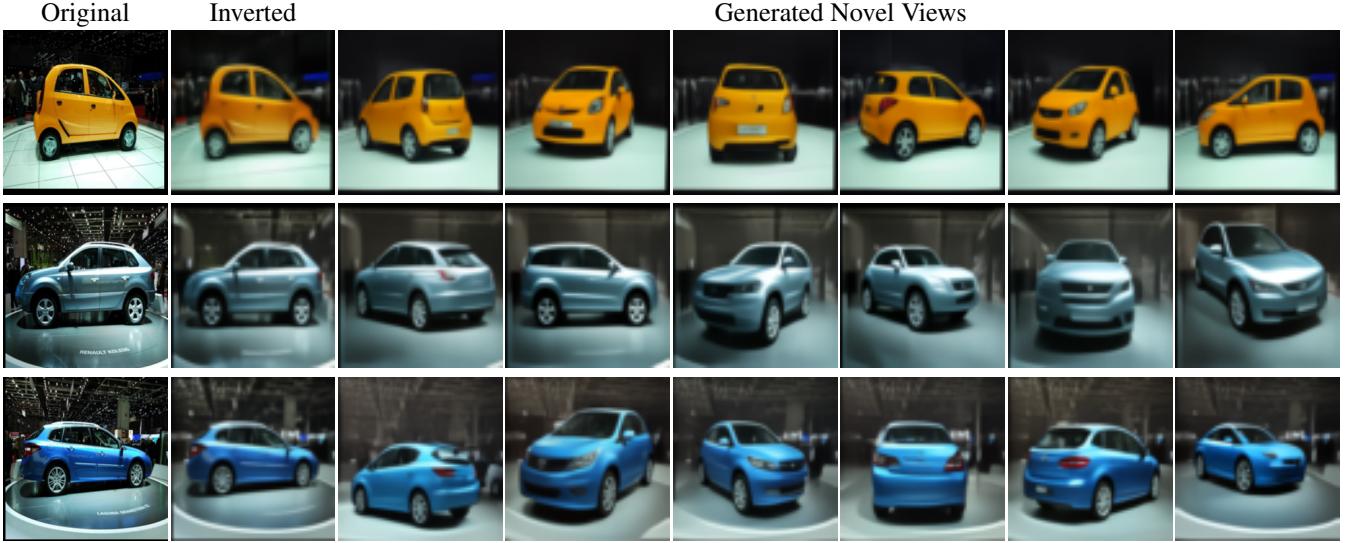


Figure 4. Additional examples of novel views generated by MIRAGE on the EPFL Dataset.

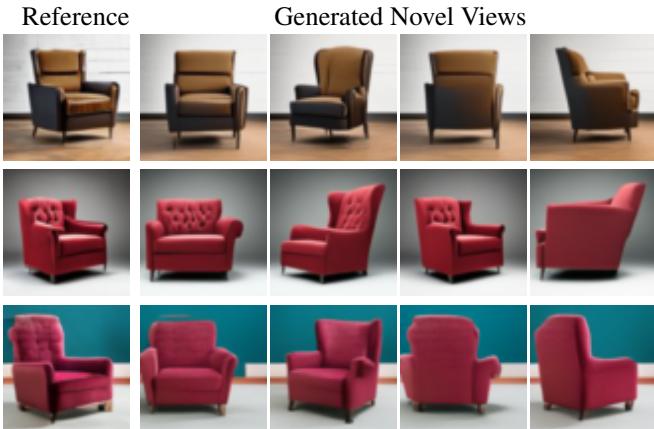


Figure 5. Additional examples of novel views generated by MIRAGE on the SD Armchairs dataset.

be the result of over-smoothing caused by HAG, since the textures in the ground truth data are already quite monotone.

We also calculated the FID for Monnier *et al.* [34]. Since their method is not generative we took the CompCars dataset and reconstructed randomly selected 30K cars and then generated 4 novel views from 15K cars from this set. Since their generated views have a white background we utilized SAM [27] to remove the background from the ground truth data. We also applied the same procedure to our generated dataset. From Table 2, we observe that the FID for Monnier *et al.* [34] is much higher compared to the our masked version. This can also be observed visually in

Figure 8. This is expected as their method is not supposed to generate high quality textures because of the cross consistency loss between similar instances.

In Figures 6 and 5 we show visual results for both generated novel views for car and armchair datasets starting from the normal noise. In Figure 4 we show the novel views for some cars in the EPFL dataset [38], which is out-of-distribution for the model trained on CompCars. For this, we first take the test image and then find its pose label using the procedure explained in section 3.1. Once we do the inversion, we generate novel views as described in section 3.

**3D Reconstruction.** In Figure 7 we showcase the obtained 3D model from novel views generated for the car from the EPFL dataset. Once we obtained novel views from our method we run NeRS [64] on a sparse set of novel views. Here we have to we initialize the camera poses manually. We use SAM [27] to obtain the masks since NeRS can work only with mask supervision. In the absence of camera poses it was difficult to assign initial poses to the generated images. For this reason, we run only on a sparse set of generated images. In Figure 7, we show the 3D-mesh obtained by running NeRS on our generated images. We observe our novel views to be good enough for obtaining explicit 3D-Mesh. Note that our generated views are more detailed in terms of texture than the ones of Monnier *et al.* [34].

## 5. Conclusion and Limitations

This paper introduces MIRAGE, a novel pipeline designed for generating multi-view data without relying on manual annotations. Our approach outperforms prior methods in generating novel view sequences and exhibits robustness

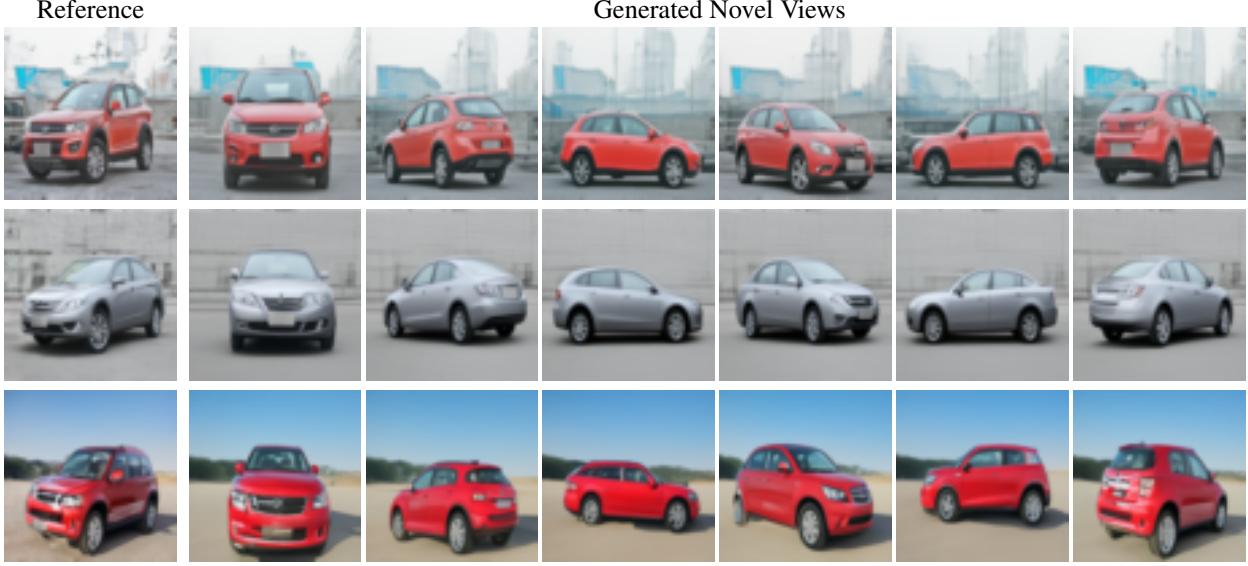


Figure 6. Additional examples of novel views generated by MIRAGE on the CompCars dataset.



Figure 7. Reconstructed 3D for a sample car in the EPFL dataset. Top row - MIRAGE (ours), bottom row - Monnier *et al.* [34]. Notice that the latter does not fit backgrounds in the model, that is why there are no backgrounds in the novel views.



Figure 8. Top 2 rows - Generated views from MIRAGE (ours). Bottom 2 row - Generated views from the reconstructed cars in the CompCars dataset using the method of Monnier *et al.* [34].

across various domains. However, a limitation is that it requires training on a dataset of single-category images. Developing an unsupervised approach for novel view synthesis on multi-category data is left for future work. We hypothesise that multi-category data can be hierarchically clustered, first to object classes (*e.g.*, by comparing DINOv2 cls tokens), and then within each class to object poses. Another limitation is the challenge in effectively distinguishing pose from orientation during clustering, especially in the case of articulated objects. Additionally, 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 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] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. *ArXiv*, abs/2311.03335, 2023. 3, 5
- [2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 1, 2
- [3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 2
- [4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 3
- [5] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 3
- [6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2
- [7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 2
- [8] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 2
- [9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3
- [10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 2
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 2
- [12] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901. 4
- [13] R. Girdhar, D.F. Fouhey, M. Rodriguez, and A. Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 3
- [14] Shubham Goel, Angjoo Kanazawa, , and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 3
- [15] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 2
- [16] Philipp Henzler, Niloy J. Mitra, and Tobias Ritschel. Escaping plato’s cave: 3d shape from adversarial rendering. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6
- [18] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 5
- [19] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 2, 3, 4
- [20] Susung Hong, Junyoung Seo, Sunghwan Hong, Heeseong Shin, and Seungryong Kim. Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023. 3
- [21] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 4
- [22] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3
- [23] Animesh Karnewar, Niloy J. Mitra, Andrea Vedaldi, and David Novotny. Holofusion: Towards photo-realistic 3d generative modeling, 2023. 3
- [24] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy Mitra. Holodiffusion: Training a 3D diffusion model using 2D images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 3
- [25] Levon Khachatryan, Andranik Moysisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 2, 3, 5
- [26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 7
- [28] 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 *Proc. IEEE*

- Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 11
- [29] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3
- [30] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 2
- [31] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 2
- [32] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Inf. Theory*, 28:129–136, 1982. 4
- [33] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3
- [34] Tom Monnier, Matthew Fisher, Alexei A. Efros, and Mathieu Aubry. Share With Thy Neighbors: Single-View Reconstruction by Cross-Instance Consistency. In *ECCV*, 2022. 3, 5, 6, 7, 8, 11
- [35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2
- [36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3, 16
- [37] Maxime Oquab, Timothée Darzet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 4
- [38] Mustafa Ozuysal, Vincent Lepetit, and Pascal Fua. Pose estimation for category specific multiview object localization. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 7
- [39] Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, and Federico Tombari. Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4391–4401, 2023. 3
- [40] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 3
- [41] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. In *ICML*, 2021. 2
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 5
- [43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 3, 4
- [44] David A. Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77:125–141, 2008. 4
- [45] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinrong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2
- [46] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3
- [47] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *ArXiv*, abs/2310.15110, 2023. 2
- [48] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2023. 3
- [49] Alessandro Simoni, Stefano Pini, Roberto Vezzani, and Rita Cucchiara. Multi-category mesh reconstruction from image collections. In *2021 International Conference on 3D Vision (3DV)*, pages 1321–1330. IEEE, 2021. 3
- [50] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3d GANs. In *Advances in Neural Information Processing Systems*, 2022. 3
- [51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 4, 5, 6, 11
- [52] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. In *arXiv:2010.04595*, 2020. 2
- [53] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 3, 4
- [54] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models, 2022. 2
- [55] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, C. L. Philip Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis, 2023. 2, 3
- [56] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In

- Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 3
- [57] Yuxin Wu and Kaiming He. Group normalization. *International Journal of Computer Vision*, 128:742 – 755, 2018. 4
- [58] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 3
- [59] Jiaolong Yang and Hongdong Li. Dense, accurate optical flow estimation with piecewise parametric model. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5, 11
- [60] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion, 2023. 2
- [61] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models, 2023. 2, 3
- [62] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3
- [63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2
- [64] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 7
- [65] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *ArXiv*, abs/2010.07492, 2020. 1
- [66] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

## A. Additional Implementation Details

As stated in the main paper, during training, we use the hyperparameters outlined in DDIM [51] and follow the code from the official implementation \*. We adopt the U-Net architecture from this source \*. To make the architecture pose-conditioned, we incorporate learnable pose embeddings into the time embedding. We train our diffusion models on the CompCars [59], SD Armchairs, and SD Mugs datasets (the latter is described in the next section) for  $10^6$ ,  $5 \times 10^5$ , and  $7 \times 10^5$  iterations, respectively. All models are trained using four NVIDIA GeForce RTX 3090 GPUs. When applied, the cross-frame attention is used throughout all time steps. We will release the code upon the acceptance of our paper.

\*<https://github.com/ermongroup/ddim>

\*\*<https://github.com/filipbasara0/simple-diffusion/tree/main/model>

Dataset	Method	FID ↓
	Ours - no MV	5.228
SD Mugs	Ours - MV, w/o HAG	11.666
	Ours - MV, w/ HAG	15.487

Table 2. The FID scores for SD Mugs for various versions of MI-RAGE. MV refers to multi-view.

## B. Additional Experimental Results

In this section, we report some additional results of MI-RAGE that could not be included in the main paper due to the page number limit.

First, we show further samples of generated novel views on the CompCars dataset (see Figures 9 and 10). Notice that the model is able to generate plausible novel views from different reference poses. We observe that the greater the visibility of features in the reference image, the lower the variability in shape/textures observed in the generated views, resulting in improved consistency that aligns with our expectations.

We also show additional samples from our model trained on the SD Armchairs dataset in Figure 12. Again, regardless of the reference pose, the model is able to generate plausible novel views.

To emphasize the robustness of our pipeline to different categories, we generate another synthetic dataset from Stable Diffusion. We follow the same steps as for the SD Armchairs dataset, but this time we used the category of “a single colorful mug”. We call the resulting dataset *SD Mugs*. The qualitative results of novel view synthesis on the SD Mugs dataset is shown in Figure 13, while some quantitative data is reported in Table 2. Notice, that due to the handle, mugs have a fundamentally different topology compared to cars and armchairs. For instance, [34] are restricted to modelling objects that are homeomorphic to spheres. MI-RAGE does not have such a restriction and is able to work for diverse object topologies.

Lastly, we also show some samples from GIRAFFE 14, 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 known 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 upsample and lack of explicit 3D supervision” (as quoted from [28]).

### C. Failure Cases

We observe that the quality of the generated views highly depends on the quality of the reference pose. If the generated reference pose is not a valid image, then the generated views lack both realism and multi-view consistency. This can be clearly observed in Figure 11, where the generated novel views for two invalid reference images are both unrealistic and lack multi-view consistency.

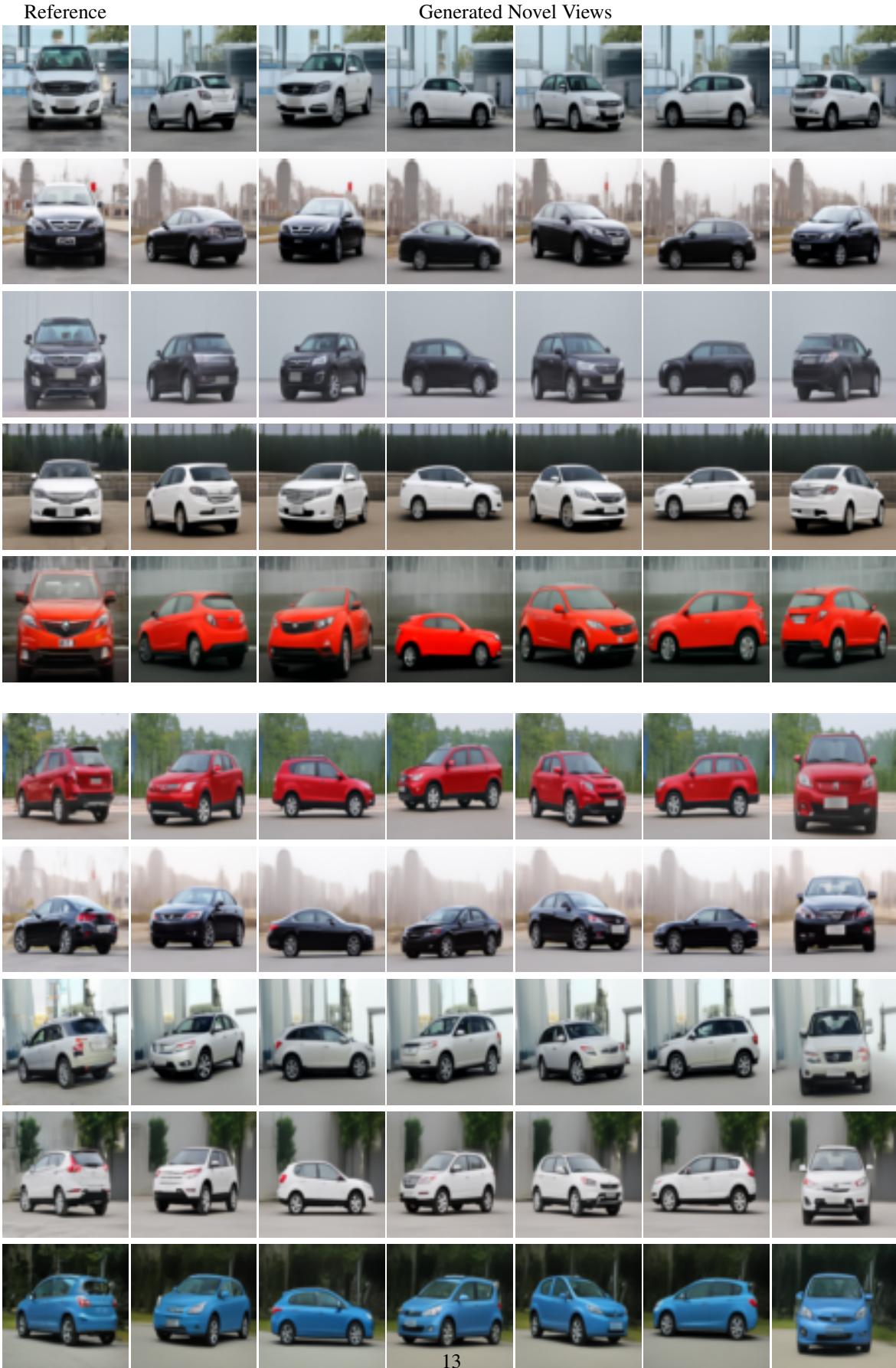


Figure 9. Generated cars with MIRAGE for different reference poses (front view and half-back view).

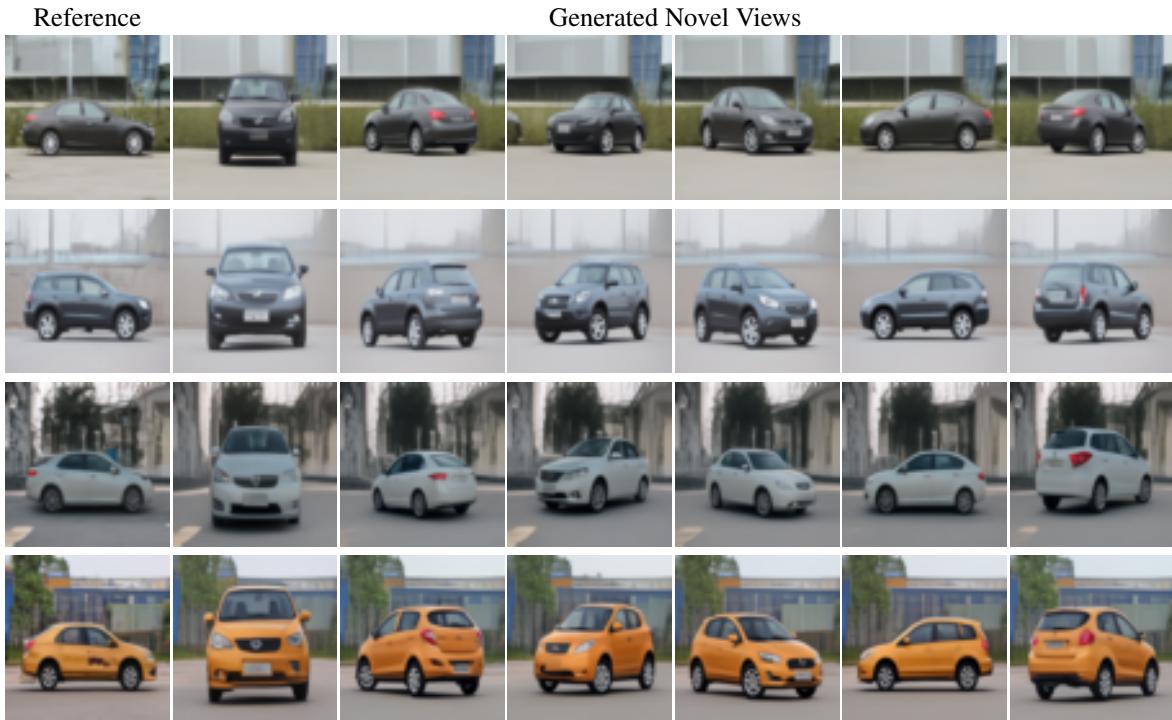


Figure 10. Generated cars with MIRAGE for the case of the side-view as the reference pose.

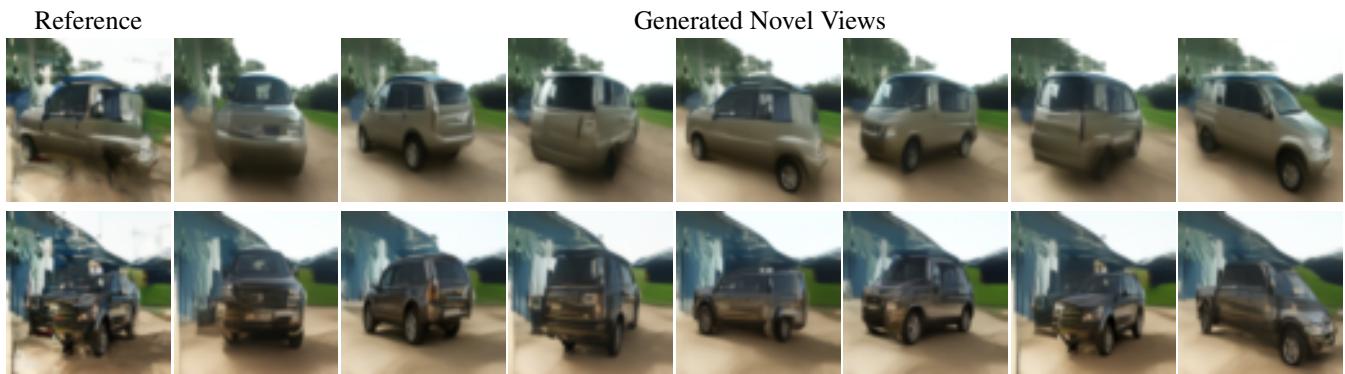


Figure 11. Failure cases for the generated cars.

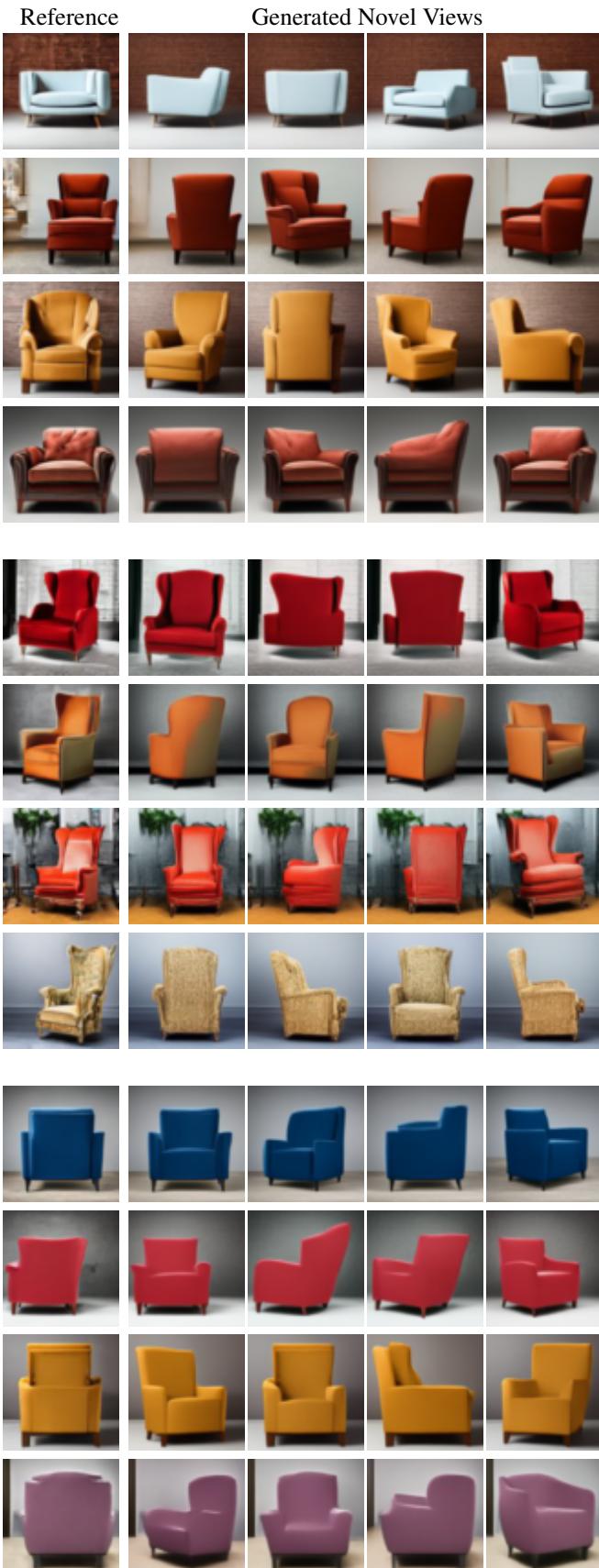


Figure 12. Additional examples of novel views generated by MiRAGE on the SD Armchairs dataset from different reference poses (front, half-front, and back views).

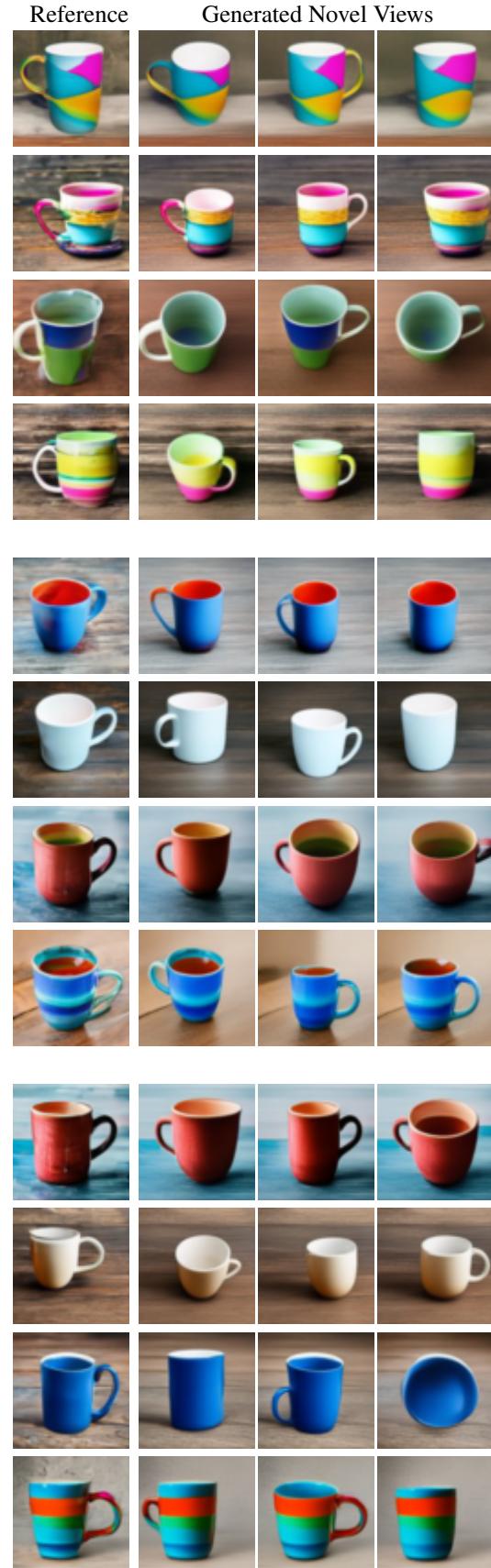


Figure 13. Additional examples of novel views generated by MiRAGE on the SD Mugs dataset from different reference poses (side views and top-side view). Note that columns do not necessarily represent the same pose cluster.

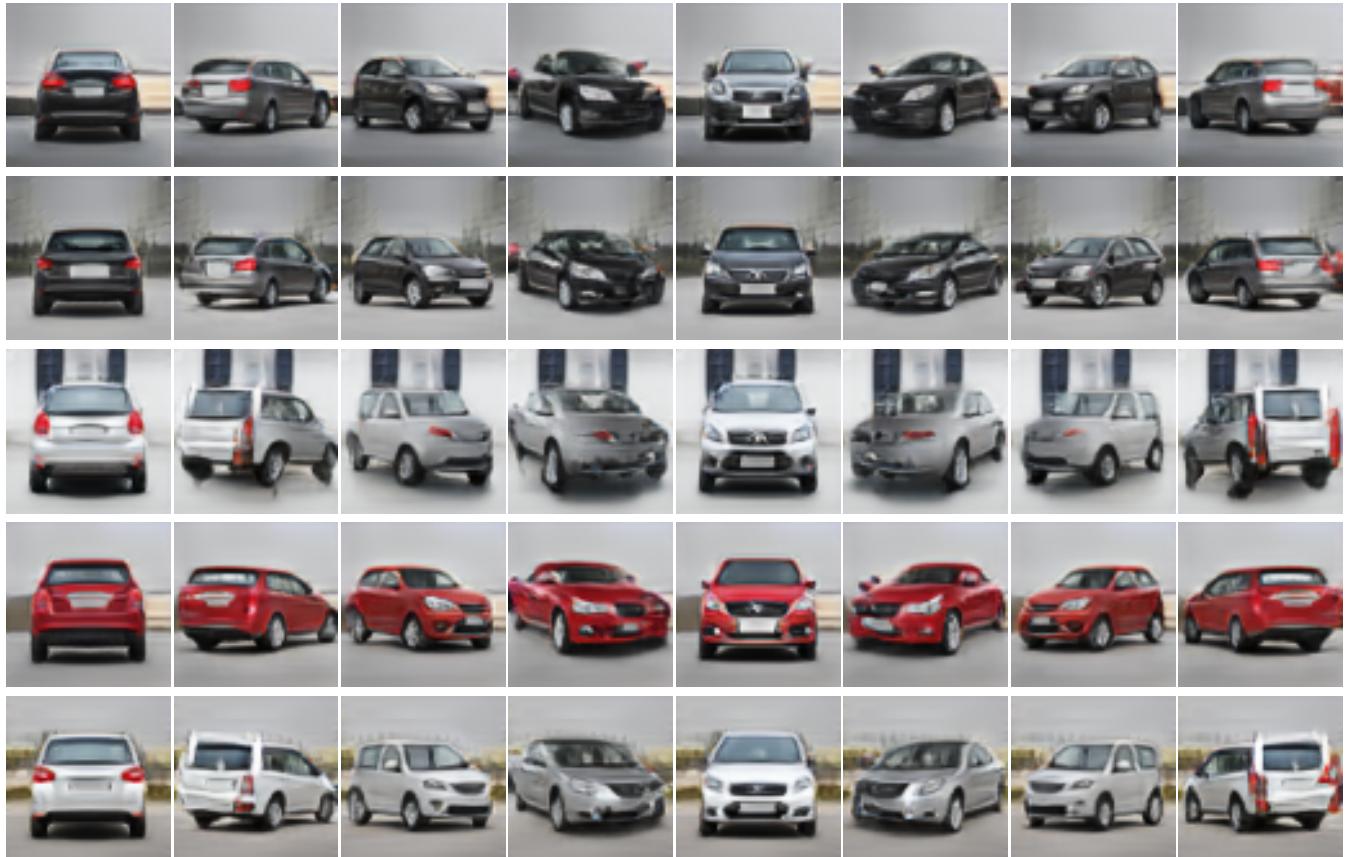


Figure 14. For comparison, 360° generated cars using GIRAFFE [36].