Enhancing 3D Fidelity of Text-to-3D using Cross-View Correspondences

Seungwook Kim[1,2] Kejie Li[2] Xueqing Deng[2] Yichun Shi[2] Minsu Cho[1] Peng Wang[2]

[1]POSTECH, South Korea [2]ByteDance, USA

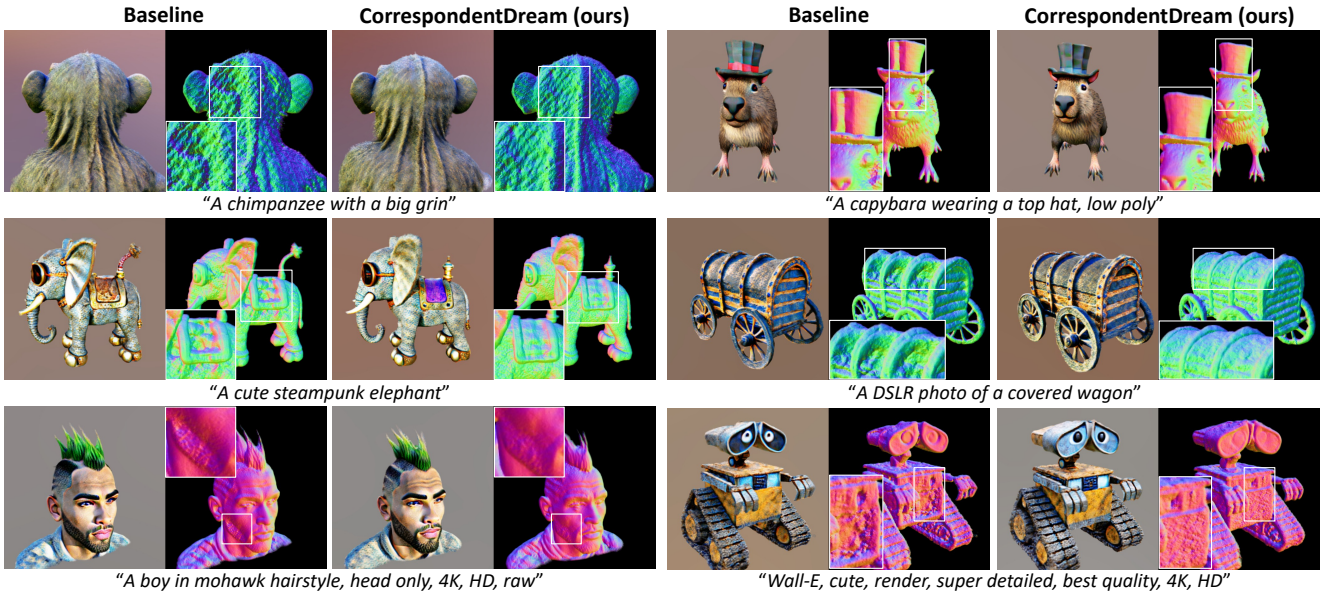

Figure 1. **Comparison between the baseline (MVDream [27]) and CorrespondentDream (ours).** Our method substantially alleviates the 3D geometric infidelity issue in zero-shot text-to-3D generation methods. Best viewed on electronics, zoom in for clearer visualization.

Abstract

Leveraging multi-view diffusion models as priors for 3D optimization have alleviated the problem of 3D consistency, e.g.*, the Janus face problem or the content drift problem, in zero-shot text-to-3D models. However, the 3D geometric fidelity of the output remains an unresolved issue; albeit the rendered 2D views are realistic, the underlying geometry may contain errors such as unreasonable concavities. In this work, we propose CorrespondentDream, an effective method to leverage annotation-free, cross-view correspondences yielded from the diffusion U-Net to provide additional 3D prior to the NeRF optimization process. We find that these correspondences are strongly consistent with human perception, and by adopting it in our loss design, we are able to produce NeRF models with geometries that are more coherent with common sense,* e.g.*, more smoothed object surface, yielding higher 3D fidelity. We demonstrate the efficacy of our approach through various comparative qualitative results and a solid user study.*

1. Introduction

Text-to-3D generation holds wide applicability in areas such as virtual reality and 3D content generation [25], which are of integral importance in the fields of gaming and media. In recent studies, leveraging 2D diffusion models as priors to optimize 3D representations, *e.g.*, NeRF [23] or NeuS [33], via Score Distillation Sampling (SDS) has shown to yield promising results and generalizability for zero-shot text-to-3D generation [18, 25].

It was subsequently observed that using a single-view 2D diffusion model as prior suffers from the lack of multi-view knowledge and 3D awareness, frequently resulting in issues including the Janus face problem or content drift [9, 27]. This problem was largely alleviated by leveraging a *multi-view* diffusion model [27] as the prior instead, which generates multiple view-consistent 2D images instead of a single 2D image to improve the multi-view consistency of the 3D output. However, even with the integration of multi-view diffusion models, the disparity in dimensionality between the 2D priors and the final 3D representation makes it insufficient to ensure the 3D geometric fidelity of the output

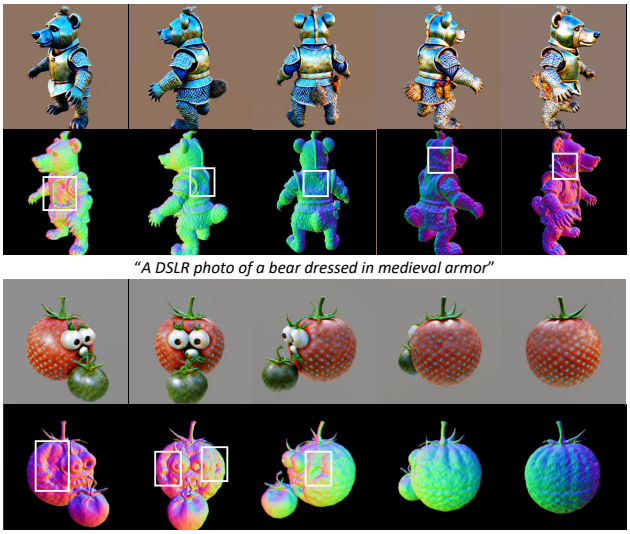

Figure 2. **Rendered 2D view and 3D normal map of MVDream [27].** While the rendered 2D views look realistic, the underlying 3D geometry lacks fidelity, with concavities or missing surfaces (highlighted in white squares).

shape, as exemplified in Fig. 2.

In this paper, we introduce CorrespondentDream, a novel method which enhances 3D fidelity in text-to-3D generation, using *cross-view correspondences* computed from the diffusion model functioning as the optimization prior. By utilizing features from upsampling layers of the diffusion U-Net, we can establish robust correspondences between multi-view images without explicit supervision or fine-tuning. Our approach hinges on the multi-view consistency of 2D features in the multi-view diffusion model, which we conjecture to be faithful to human perception.

By using the known camera parameters for NeRF-rendered views, we can reproject pixels across different views using the NeRF-rendered depth values. In the presence of 3D infidelities such as concavities or missing surfaces, the NeRF-rendered depth values will also be erroneous, reflecting the infidelities. We aim to correct these errors by aligning the NeRF reprojections with cross-view correspondences, thereby enhancing the 3D fidelity of the output by correcting the NeRF depths. The effectiveness of CorrespondentDream is validated through extensive qualitative assessments and a user study.

The contributions of our work are threefold:

- We identify that 3D infidelities remains an issue in existing zero-shot text-to-3D methods, even with improved 3D consistency via multi-view diffusion priors.
- We introduce CorrespondentDream, a novel method to incorporate *cross-view correspondences* into the 3D optimization for improved 3D fidelity.
- We demonstrate the effectiveness of our method via comparative analysis and a user study.

2. Related Work

Text-to-3D using 2D diffusion models. Based on the observation that template-based generation pipelines and 3D generative models [1, 7, 13, 34] show limited 3D generation performances due to the lack of sufficiently large-scale 3D data, 2D-lifting methods have gained interest [18, 25]. Specifically, DreamFusion [25] proposed Score Distillation Sampling (SDS) to leverage 2D diffusion models as priors to optimize 3D representations [23, 33] to facilitate zero-shot text-to-3D generation, while SJC [32] concurrently proposed a similar technique using the stable-diffusion model [26]. Subsequent studies aim to improve the output representation [2, 29], sampling schedules for optimization [10], and loss design [35] for improved quality and efficiency. However, using a single-view 2D diffusion prior is observed to suffer from multi-view inconsistency - namely the Janus face problem and content drift.

Text-to-3D using multi-view diffusion models. To alleviate the problem of multi-view inconsistency, a promising direction is to leverage improved multi-view knowledge. To this end, MVDream [27] finetunes the stable diffusion [26] model to generate multi-view images instead of a single-view image. This is facilitated by replacing the self-attention of the diffusion U-Net with multi-view attention, such that the multiple views can attend to one another for multi-view knowledge. Using multi-view diffusion as the prior to facilitate text-to-3D generation shows highly improved multi-view consistency in the rendered views.

Albeit their efficacy in addressing multi-view inconsistency between 2D rendered views, multi-view diffusion models still fall short in fully capturing the true fidelity of the underlying 3D geometry. In this work, we address this issue via integrating cross-view correspondences from the diffusion network to enforce additional geometric priors.

Establishing correspondences using diffusion models. With recent advancements in diffusion models [8, 24, 26], the potential of their representational abilities triggered many applications to visual correspondence. Unsupervised methods already exhibit competitive performances, relying on iterative refinement of features [6] or an additional feature extractor [36] for improved performance. The performance gains were notably higher in a strongly-supervised training scheme, either via aggregating the multi-scale features from multiple timesteps [22], or by optimizing pair-specific prompts for better matchable features [17].

In this work, we take inspiration from DIFT [30] to yield cross-view correspondences using diffusion features.

Leveraging correspondences for NeRF optimization. NeRF [23] encodes 3D scenes with a MLP, which can eventually be used for 2D view rendering. In recent studies, photometric loss alone proved to be insufficient to train a NeRF model under challenging constraints, *e.g.*, sparse in-

put views [3, 11, 31] or erroneous camera poses [12, 31]. Using off-the-shelf image matching models as additional priors has shown promising results under sparse viewpoints or noisy poses [12, 31]. SCNeRF [12] aims to minimize the projected ray distance of off-the-shelf correspondences, while SPARF [31] proposes to minimize the reprojection error between the off-the-shelf matches and the reprojected matches obtained using NeRF depths.

In contrast, we compute correspondences between NeRF-rendered views, instead of ground truth images. Furthermore, we do not rely on off-the-shelf matching methods, but compute annotation-free cross-view correspondences from diffusion features to provide additional geometric priors in optimizing NeRF for improved 3D fidelity.

3. Preliminary: Text-to-3D using Diffusion

DreamFusion [25] introduced Score Distillation Sampling (SDS), which facilitates the optimization of differentiable image parameterizations (DIP) by using diffusion models to compute gradients in the form of:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, x = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi(z_t; y, t) - \nabla_\theta x\right)\right]. \tag{1}$$

In this formulation, θ denotes the parameters of the DIP, ϕ represents the parameters of the diffusion model, and x signifies the image rendered by the DIP through the function g. The term $w(t)$ is a weighting function dependent on the sampled timestep t. The variable ϵ stands for the noise vector, and z_t is the noisy image at timestep t. The expectation $\mathbb{E}$ is taken over both t and ϵ, with y being the conditioning variable, such as a text prompt. This approach allows a 2D diffusion model to act as a 'frozen critic', predicting image-space modifications to optimize the DIP. Common DIPs include 3D volumetric representations such as NeRF [23] or NeuS [33], thus enabling zero-shot text-to-3D generation.

However, a challenge arises from the lack of 3D consistency across rendered views due to the absence of integrated multi-view knowledge. Research has shown promising results in addressing this challenge by introducing multi-view attention within the diffusion U-Nets to facilitate the training of multi-view diffusion models. These models yield consistent multi-view color images, substantially improving 3D coherency [20, 27, 28]. With a multi-view diffusion model at disposal, MVDream [27] defines the multi-view diffusion loss for supervising the 3D volume as:

$$\mathcal{L}_{\text{SDS}}(\phi, \{x_i = g(\phi, c_i)\}_{i=1}^N) = \mathbb{E}_{t,c,\epsilon}\left[\sum_{i=1}^N \|x_i - \hat{x}_{0,i}\|^2\right] \tag{2}$$

Here, c_i denotes the camera pose for the i-th view, x_i is the image rendered from the 3D volume for the i-th view, and $\hat{x}_{0,i}$ is the corresponding image generated by the diffusion model. This deviates from the original SDS formulation (Eq. (1)); MVDream proposes that Eq. (2) is equivalent to Eq. (1) with $w(t)$ and shows to perform similarly as well, but using Eq. (2) further enables the usage of CFG rescale trick [19] to mitigate color saturation. The improved 3D consistency afforded by these multi-view diffusion models ensures that the 3D volume's rendered views adhere to the same level of coherence, leading to substantial improvements in the stability and quality of text-to-3D generation.

4. Method

Motivation and overview. Employing multi-view diffusion as a prior for 3D generation enhances the consistency of NeRF-rendered views, mitigating common issues such as the Janus face problem or content drift. However, these priors are still confined to 2D space, which often results in errors in the geometric fidelity of the 3D output. While 2D rendered views may appear realistic, the 3D geometry can be flawed, exhibiting issues such as unnatural concavities or missing surfaces, as shown in Fig. 2.

We introduce CorrespondentDream, a novel method designed to improve the 3D fidelity of zero-shot text-to-3D outputs, using cross-view correspondences derived from the multi-view diffusion prior. Our approach involves optimizing a NeRF model through both SDS and cross-view correspondence losses. As the SDS loss adopts the conventional form as seen in Eq. (2), we detail the correspondence loss in this section. We generate two adjacent sets of NeRF-rendered views with minimally separated camera positions in azimuth (Sec. 4.1), extract 2D features from the diffusion model (Sec. 4.2), compute cross-view correspondences between adjacent rendered views (Sec. 4.3), and use them to correct NeRF geometry via the cross-view correspondence loss (Sec. 4.4). The NeRF optimization using the SDS and cross-view correspondence losses are detailed in Sec. 4.5. Fig. 3 illustrates an overview of CorrespondentDream.

4.1. Adjacent multi-view NeRF rendering

Our approach utilizes a multi-view diffusion model ϵ_θ which can concurrently generate N images $\{x_t^{(i)}\}_{i=1}^N$, where each image is associated with a distinct viewpoint derived from the camera parameters and the given textual prompt y. These images represent a range of equispaced azimuth angles, capturing different perspectives of the same scene. To effectively optimize the NeRF model ϕ, we would follow Eq. (2) to render N views $\{g(\phi, c_i)\}_{i=1}^N$, where c_i defines the camera parameters corresponding to the i-th view, and g is the NeRF rendering function dependent on the parameters of ϕ. Through this process, the NeRF model is supervised to produce images that are consistent with the specified perspectives of c_i, aligning the NeRF-rendered views with the diffusion model's predictions.

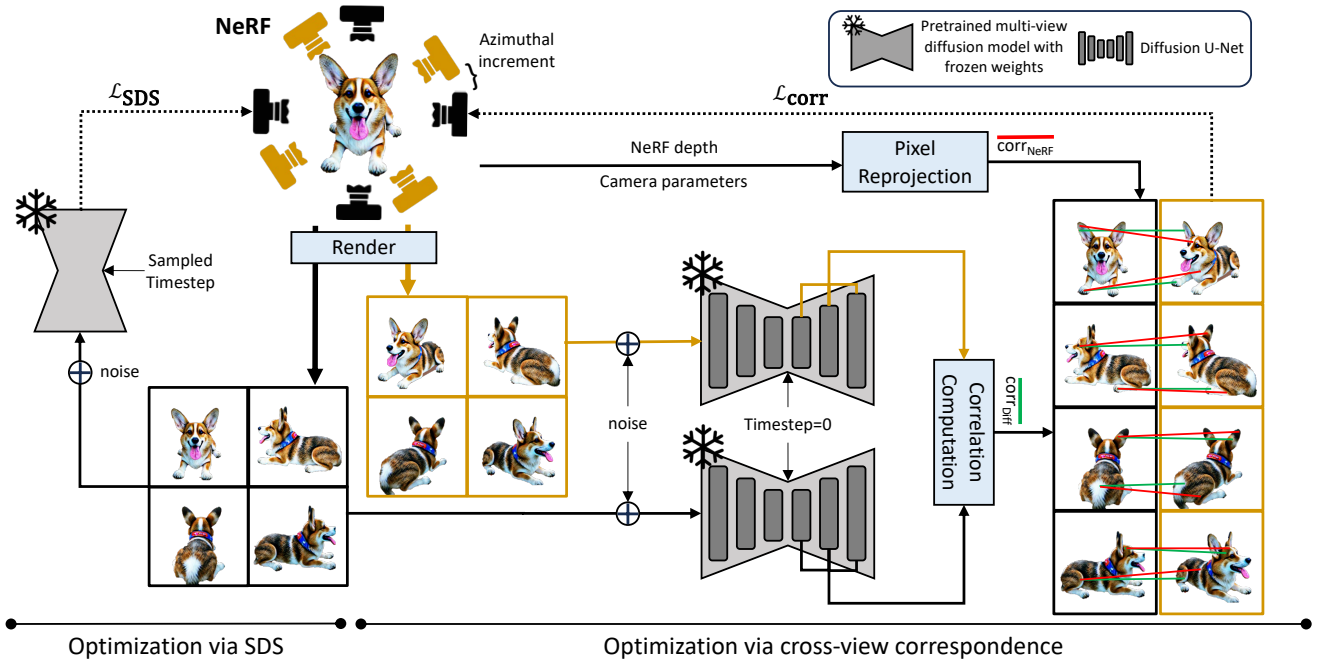

Figure 3. **Overview of CorrespondentDream**. We employ NeRF [23] for 3D representation, optimized alternately using the SDS loss ($\mathcal{L}_{\text{SDS}}$) and cross-view correspondence loss ($\mathcal{L}_{\text{corr}}$). The $\mathcal{L}_{\text{SDS}}$ is based on the multi-view formulation from Eq. (2) in MVDream [27]. To compute $\mathcal{L}_{\text{corr}}$, we render two adjacent view sets from NeRF with identical noise, inputting them into a frozen pre-trained multi-view diffusion model. We then extract multi-layer features from the diffusion U-Net's upsampling layers to establish correspondences ($\text{corr}_{\text{diff}}$) between each view pair. Utilizing ground-truth camera parameters and NeRF-rendered depth, we reproject pixels to obtain $\text{corr}_{\text{NeRF}}$. By minimizing the discrepancy between $\text{corr}_{\text{NeRF}}$ and $\text{corr}_{\text{diff}}$, the pseudo ground-truth, we correct NeRF's 3D infidelities in the NeRF depths.

Due to the large azimuthal distances between each of the N views, there is limited viewpoint overlap between adjacent views, making direct correspondence computation challenging and prone to error. To address this, we render two interlinked sets of N views, V_1 and V_2, ensuring that each view in V_1 has an adjacent view in V_2, thereby minimizing azimuthal separation. The azimuth angles for the two sets are articulated as $\{\alpha_i\}_{i=1}^{N}$ and $\{\beta_i\}_{i=1}^{N}$, where β_i is defined as $\alpha_i + \Delta\alpha$, with $\Delta\alpha$ being a small, predetermined angular increment. This approach simplifies the computation of correspondences by providing more overlapping fields of view, thus ensuring a robust set of correspondences for subsequent optimization processes.

4.2. Annotation-free feature extraction

We take advantage of the U-Net architecture within our multi-view diffusion model, which is adept at generating N synchronized views. In the optimization of 3D representations for text-to-3D generation, we add Gaussian noise η to the NeRF-rendered views, modulated by a timestep t, creating noisy images $\tilde{v}_t^{(i)} = v_t^{(i)} + \sqrt{\alpha_t}\eta$, with η distributed as $\mathcal{N}(0, \mathbf{I})$, and α_t being a variance schedule function of the timestep t, which controls the noise level. The diffusion model ϵ_θ then predicts the noise component as:

$$\hat{\eta}^{(i)} = \epsilon_\theta(\tilde{v}_t^{(i)}; y, c_i, t)$$

During this predictive step, we extract intermediate features $\{f_l^{(i)}\}$ from the U-Net's upsampling layers l. We build on existing studies that demonstrate the robustness of multi-layer features [14, 22] to extract intermediate features across multiple layers. These features are expressed as:

$$f_l^{(i)} = U_l(\tilde{v}_t^{(i)}; \theta_l) \tag{3}$$

where U_l is the upsampling function at layer l, and θ_l are the learned parameters specific to that layer. This process yields a comprehensive set of features without additional training or explicit feature extraction algorithms. Previous studies [17, 30] show that these diffusion U-Net features are surprisingly informative and discriminative, enabling the establishment of robust image correspondences.

4.3. Cross-view correspondence computation

After obtaining the multi-view features for $2N$ views, $\{f_l^{(i)}\}$ and $\{f_l^{(i+N)}\}$ for $i = 1$ to N, we compute the correspondences between each pair of adjacent views, yielding N sets of adjacent-view correspondences. The feature maps extracted from the diffusion U-Net possess varying spatial dimensions across different layers, and are interpolated to a common resolution $H' \times W'$, as follows:

$$f_l^{'(i)} = \mathcal{B}(f_l^{(i)}, H', W'), \quad f_l^{'(i+N)} = \mathcal{B}(f_l^{(i+N)}, H', W') \tag{4}$$

where $\mathcal{B}$ represents the bilinear interpolation function.

Prior to computing the correlation map, the feature maps are normalized to ensure comparability. The correlation map $C_l^{(i)}$ at each feature level l is computed to encapsulate pairwise similarity across all spatial positions, resulting in a 4D tensor with dimensions $H' \times W' \times H' \times W'$. The element $C_l^{(i)}(p, q, r, s)$ represents the L2 distance between the vectors at positions (p, q) and (r, s):

$$C_l^{(i)}(p,q,r,s) = \frac{f_l^{'(i)}(p,q) \cdot f_l^{'(i+N)}(r,s)}{\left\| f_l^{'(i)}(p,q) \right\|_2 \left\| f_l^{'(i+N)}(r,s) \right\|_2} \quad (5)$$

Subsequently, we aggregate the correlation maps from all feature levels to form the cumulative correspondence map $\mathcal{C}^{(i)}$ for each view i, defined by the sum:

$$\mathcal{C}^{(i)} = \sum_l C_l^{(i)} \quad (6)$$

This 4D correlation map $\mathcal{C}^{(i)}$ integrates the feature-level similarities into a singular comprehensive map [15].

For each spatial location (p, q), correspondences are determined by identifying the position with the highest value in $\mathcal{C}^{(i)}$, signifying the nearest neighbor:

$$\text{corr}(p,q) = \arg\max_{r,s} \mathcal{C}^{(i)}(p,q,r,s) \quad (7)$$

where $\text{corr}(p, q)$ designates the corresponding spatial location in the adjacent view for the point at (p, q). This dense correspondence field between each pair of adjacent views serves as a 3D geometric prior, which is instrumental in supervising the NeRF model for improved 3D fidelity.

As we know the ground-truth camera parameters for each rendered view, we can filter out implausible correspondences, adhering to constraints like the epipolar constraint. We guide the readers to the supplementary for the details of correspondence post-processing.

4.4. Cross-view correspondence loss

Having established N sets of correspondences between adjacent NeRF-rendered image pairs, we leverage the depth information provided by the NeRF rendering process alongside the known camera parameters to reproject points from one view to the corresponding points in the adjacent view through a reprojection function denoted as π. For each pair of adjacent images, we now possess two distinct sets of correspondences: one derived from the diffusion features, $\text{corr}_{\text{diff}}$, and the other obtained via reprojection using camera parameters and NeRF-rendered depths, $\text{corr}_{\text{NeRF}}$. The reprojection function is defined as:

$$\text{corr}_{\text{NeRF}}(p) = \pi(\text{depth}_\phi(p), c, p) \quad (8)$$

where $\text{depth}_\phi(p)$ is the depth value at pixel p, and c represents the camera parameters.

Our assumption posits that diffusion features are both informative and discriminative, yielding correspondences that not only associate semantically similar features but also adhere to geometric consistency, aligning with human perceptual reasoning. To enforce this assumption, we take inspiration from SPARF [31] to formulate a cross-view correspondence loss, which penalizes the NeRF-reprojected correspondences when they diverge from the diffusion feature correspondences, *i.e.*, incoherent to common sense:

$$\mathcal{L}_{\text{corr}} = \sum_p \omega(p) \cdot \text{Huber}(\text{corr}_{\text{diff}}(p), \text{corr}_{\text{NeRF}}(p)) \quad (9)$$

Here, $\text{Huber}(\cdot)$ represents the Huber loss function [5], and $\omega(p)$ is a weighting factor proportional to the similarity value at position p in the correlation map, enhancing the influence of high-confidence correspondences. This loss function serves to align the NeRF model's depth predictions with the geometrically and semantically robust correspondences derived from diffusion features, thereby correcting infidelities in the NeRF-rendered depths and enhancing the model's coherence to common sense.

4.5. NeRF optimization

In optimizing the NeRF model, we consider two distinct objectives: $\mathcal{L}_{\text{SDS}}$, which ensures that NeRF-rendered views are consistent with the pre-trained diffusion model, and $\mathcal{L}_{\text{corr}}$, which improves the 3D fidelity of the NeRF-inferred geometry. Given the potential for these objectives to conflict—wherein the 3D geometry updates from $\mathcal{L}_{\text{corr}}$ may not align with updates steered by $\mathcal{L}_{\text{SDS}}$—we employ an alternating optimization strategy to mitigate conflict between the objectives. This strategy is particularly pertinent as features conducive to accurate correspondence are typically extracted at lower timesteps (e.g., $t = 0$), while $\mathcal{L}_{\text{SDS}}$ benefits from a wide range of sampled timesteps during the 3D optimization process [25, 27].

We define the total number of optimization iterations as T, with a predefined range $[t_{\text{start}}, t_{\text{end}}]$ within which $\mathcal{L}_{\text{corr}}$ is active. The SDS loss is applied to NeRF at every iteration by default. However, for iterations t such that $t_{\text{start}} \leq t \leq t_{\text{end}}$ and t is even ($t\%2 = 0$), we alternate to apply $\mathcal{L}_{\text{corr}}$ without $\mathcal{L}_{\text{SDS}}$. This alternating approach leverages the strengths of both losses, facilitating a balanced optimization that enhances both the visual coherence and 3D geometric fidelity of the NeRF-rendered scenes.

5. Experiment

5.1. Implementation details

We implement CorrespondentDream on top of the open-sourced multi-view diffusion model MVDream [27], which implements a zero-shot text-to-3D pipeline on top of the threestudio [4] library. We use final image dimensions of

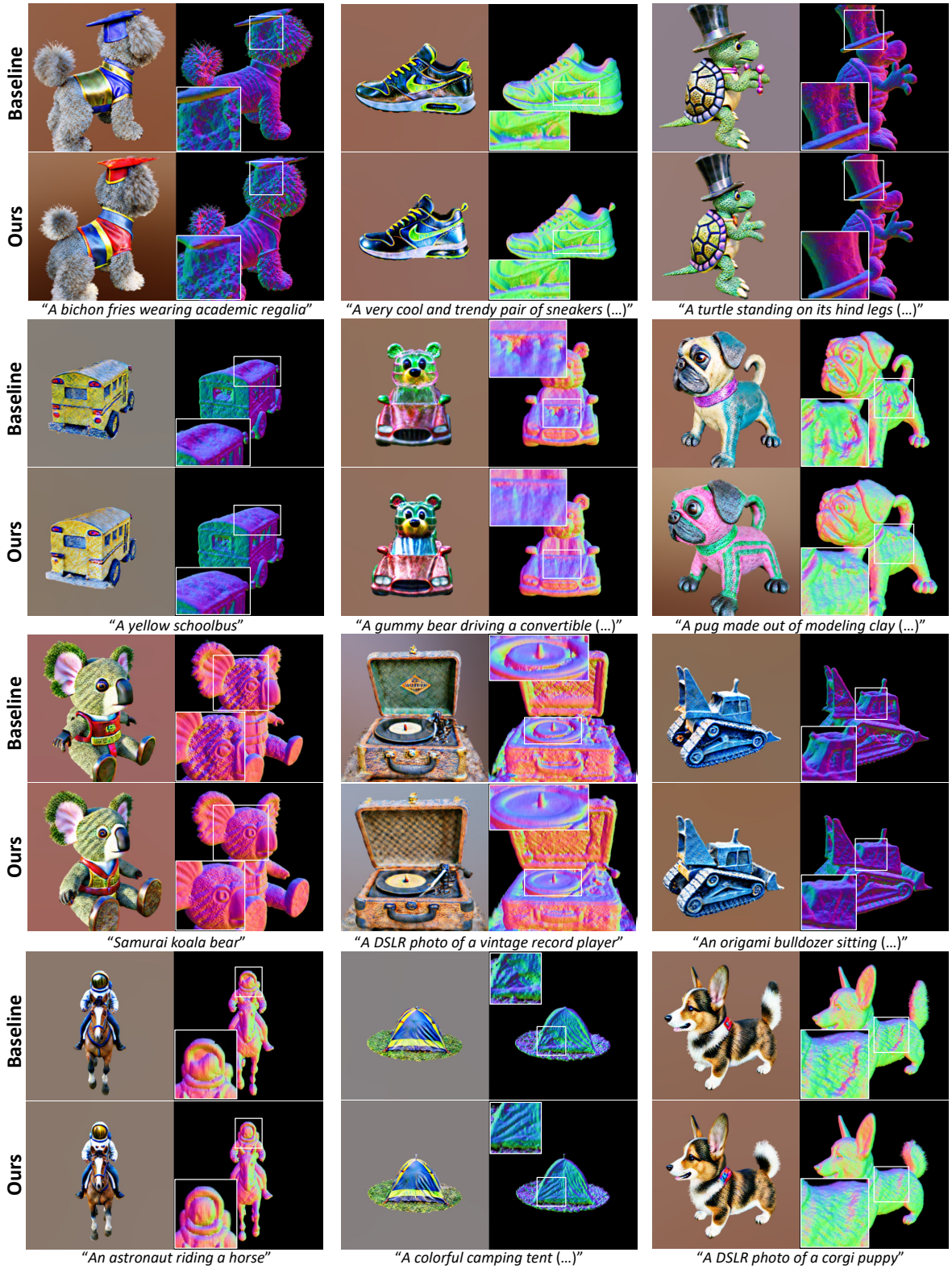

Figure 4. **Qualitative results of our CorrespondentDream across various prompts**. It can be seen that CorrespondentDream yields substantially improved 3D fidelity across various prompts. The 3D infidelities from the baseline (MVDream [27]) are highlighted and zoomed in white squares. Best viewed on electronics, zoom in for better visualization.

128×128 for NeRF-rendered views for improved latency and memory overhead[1]. We noticed that the quality of text-to-3D using MVDream is maintained at image dimensions of 128×128; and in the presence of 3D infidelities, the errors are also consistent at these image resolutions. We illustrate qualitative evidence for this in the supplementary.

For our NeRF, we use the implicit-volume implementation in the Nerfacc library [16]. We use $t_{\text{start}} = 3000$ and $t_{\text{end}} = 7000$, leading to 2000 iterations of correspondence loss supervision without SDS supervision. To compensate for this, we optimize our NeRF model for a total number of iterations $T = 12000$[2]. We optimize the NeRF model using an AdamW optimizer [21] with a constant learning rate of 0.01. L_{SDS} and L_{corr} are weighted at 1.0 and 1,000 respectively. For adjacent multi-view NeRF rendering, we uniformly sample from $[10°, 30°]$ for $\Delta\alpha$, and ensure that the adjacent views are added with the same noise, modulated with $t = 0$[3]. We use the output feature maps from the 6th and 9th upsampling layers of the UNet to compute the correlation map. The NeRF optimization with normal rendering takes about 2 hours on a Tesla V100 GPU.

5.2. Qualitative results

We present comparative qualitative results between CorrespondentDream and MVDream [27]. Other 2D lifting methods that rely on single-view diffusion model priors [18, 25] suffer from unresolved multi-view consistency issues, *i.e.*, the Janus face and the content drift problems. This overwhelms the 3D infidelities, making in inappropriate to qualitatively compare against such methods (**??**). The results are shown in Fig. 4, where it can be seen that CorrespondentDream can visibly remove 3D infidelities *i.e.*, concavities or missing surfaces, across various prompts, improving the 3D fidelity of NeRF-rendered geometry.

As our proposed cross-view correspondence affects the NeRF model directly via the NeRF-rendered depths, it can be seen that the output appearance *i.e.*, size, color or the overall appearance, of the output itself may also differ compared to using the SDS loss alone. For example, the 3D outputs of the prompt "*A bichon fries wearing academic regalia*" are wearing differently coloured regalias, and the 3D outputs of the prompt "*A pug made out of modeling clay*" exhibit different colours as well. Nonetheless, the output when using our method is still coherent with the input prompt, but with improved 3D fidelity.

[1]Stable Diffusion v2.1 base model [26] generates images at resolutions of 512×512. MVDream [27] uses a reduced image size of 256×256 when finetuning the Stable Diffusion model for multi-view image generation, and also when rendering NerF-rendered views for text-to-3D.

[2]which is 2000 higher than the number of iterations used in MVDream.

[3]Inspired from existing work on diffusion-based image matching [17, 30] which show that $t = 0$ gives the most robust features.

5.3. Comparative analysis

In this section, we perform analytical experiments to qualitatively evidence the design choices of our approach.

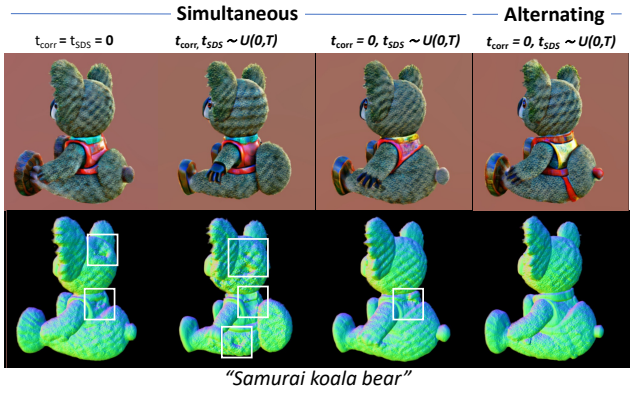

Figure 5. **Analysis of alternating supervision.** Noticeable 3D inaccuracies are marked with white squares. Our alternating supervision approach demonstrates superior qualitative outcomes.

Analysis on alternating supervision of $\mathcal{L}_{\text{SDS}}$ and $\mathcal{L}_{\text{corr}}$. We compare the scheme of supervising NeRF with alternating $\mathcal{L}_{\text{SDS}}$ and $\mathcal{L}_{\text{corr}}$ to non-alternating (simultaneous) alternatives. Under the simultaneous setting, we either (1) fix the timestep t to be always randomly sampled (as done for $\mathcal{L}_{\text{SDS}}$), or (2) always set at $t = 0$ (as done for $\mathcal{L}_{\text{corr}}$), or (3) randomly sample the timestep t for $\mathcal{L}_{\text{SDS}}$, and use $t = 0$ for $\mathcal{L}_{\text{corr}}$ when modulating the diffusion model. Note that the (3) results in increased computation costs as the consequence of multiple forwards of the diffusion model for the same input. We illustrate the results in Fig. 5, where it can be seen that our alternating scheme yields the best results.

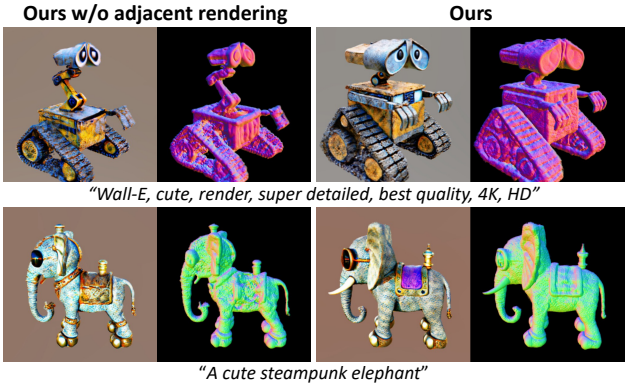

Figure 6. **Ablation of adjacent multi-view rendering.** Without adjacent rendering, we establish correspondences between adjacent views within only a single set of rendered views. This results in a very small overlapping region, leading to erroneous cross-view correspondence, and consequently worsened output.

Ablation on adjacent multi-view rendering. We perform an ablation on adjacent multi-view rendering, and show the results in Fig. 6. It clearly shows that our current scheme

of adjacent multi-view rendering yields much better results. As explained in Sec. 4.1, the azimuthal distance between adjacent views within a single set of multi-view renderings would be too large, *i.e.*, lower overlap region between viewpoints, making it challenging to establish dense, robust correspondences for appropriate supervision.

5.4. User study

Method	User preference %
MVDream [27]	30.4
CorrespondentDream (ours)	**69.6**

Table 1. **User study.** Users were asked to pick their preference based on perceived 3D fidelity and overall quality. Our method was selected more than twice compared to the baseline [27].

Due to the absence of ground-truth 3D scenes corresponding to text prompts, it is difficult to conduct a quantitative evaluation on the 3D fidelity of the text-to-3D outputs. Instead, we perform a user study on generated models from 40 non-cherry-picked prompts, where each user is asked to select the preferred 3D model in terms of the 3D fidelity *i.e.*, coherence to the 2D images, and the overall quality of the output. 767 responses from 25 participants were collected, and the results are shown in Table 1. It can be seen that CorrespondentDream was selected to be more favorable more than twice compared to MVDream alone.

5.5. Drawbacks and failure cases.

Figure 7. **Varying iterations.** Even with the same number of iterations as the baseline [27], *i.e.*, lower number of effective SDS optimization steps, CorrespondentDream still improves the 3D fidelity of the output while achieving high-quality colour and texture.

The main drawback of our method is the increase in optimization iterations due to the alternating supervision. However, we note that we used a larger number of iterations to keep the number of SDS-supervised iterations the same with the baseline for a fair comparison. Figure 7 shows the result when we use the same number of iterations as the baseline. CorrespondentDream still shows high-quality colour and texture albeit improved fidelity[4].

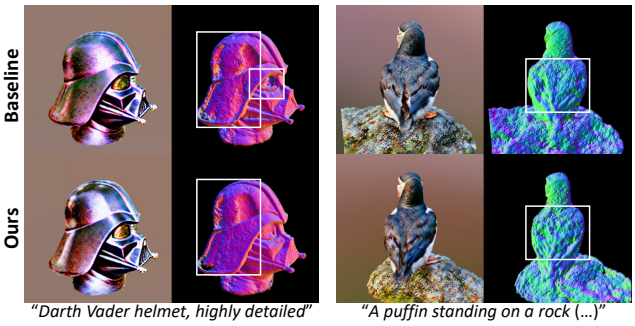

Figure 8. **Failure cases.** In the presence of shiny homogeneous surfaces (helmet, left), or many repetitive patterns (feathers, right), our method occasionally falls short at correcting the 3D infidelity.

CorrespondentDream shows to often fail in cases where there are shiny homogeneous surfaces or repeated patterns within the image as shown in Figure 8. We conjecture this is because such cases may be challenging for diffusion features to yield robust and dense correspondences between the rendered views, being unable to provide sufficiently informative 3D prior during the NeRF optimization.

6. Conclusion

We have presented CorrespondentDream, a novel method that leverages annotation-free, cross-view correspondences computed from diffusion features to additionally supervise the 3D representation in zero-shot text-to-3D models for improved 3D fidelity. By formulating the cross-view correspondence loss computed using the NeRF-reprojected pixels and the cross-view correspondences, we can correct the geometric flaws in NeRF depths caused by the ambiguities that cannot be handled by multi-view 2D image priors alone. Notably, this does not require any additional explicit priors or off-the-shelf modules. We demonstrate the efficacy of our approach via qualitative results and a user study on a large collection of varying text prompts. Our work aims to shed light onto the neglected issue of 3D geometric fidelity of diffusion-guided text-to-3D models, paving the way for enhanced applicability to practical scenarios.

Acknowledgement. This work was done while Seungwook Kim was an intern at ByteDance. Seungwook Kim was supported by the Hyundai-Motor Chung Mong-koo Foundation. This work was also supported by IITP grants (No.2021-0-02068: AI Innovation Hub (90%), No.2019-0-01906: Artificial Intelligence Graduate School Program at POSTECH (10%)) funded by the Korea government (MSIT).

[4]This *may* result in a different appearance depending on the prompt, as the effective number of iterations for SDS optimization is decreased.

References

[1] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Genvs: Generative novel view synthesis with 3d-aware diffusion models, 2023. 2

[2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2

[3] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 3

[4] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 5

[5] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009. 5

[6] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *arXiv preprint arxiv:2305.15581*, 2023. 2

[7] Paul Henderson, Vagia Tsiminaki, and Christoph H Lampert. Leveraging 2d data to learn textured 3d mesh generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7498–7507, 2020. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[9] Susung Hong, Donghoon Ahn, and Seungryong Kim. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. *arXiv preprint arXiv:2303.15413*, 2023. 1

[10] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023. 2

[11] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 3

[12] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 3

[13] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18423–18433, 2023. 2

[14] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8707, 2022. 4

[15] Seungwook Kim, Juhong Min, and Minsu Cho. Efficient semantic matching with hypercolumn correlation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 139–148, 2024. 5

[16] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: Efficient sampling accelerates nerfs. *arXiv preprint arXiv:2305.04966*, 2023. 7

[17] Xinghui Li, Jingyi Lu, Kai Han, and Victor Prisacariu. Sd4match: Learning to prompt stable diffusion model for semantic matching. *arXiv preprint arXiv:2310.17569*, 2023. 2, 4, 7

[18] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1, 2, 7

[19] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 3

[20] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 7

[22] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023. 2, 4

[23] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3, 4

[24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[25] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 5, 7

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 7

[27] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d gen-

eration. *arXiv preprint arXiv:2308.16512*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[28] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. *arXiv e-prints*, pages arXiv–2306, 2023. 3

[29] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 2

[30] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2, 4, 7

[31] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 3, 5

[32] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[33] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34:27171–27183, 2021. 1, 2, 3

[34] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 2

[35] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 2

[36] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arxiv:2305.15347*, 2023. 2