Aerial Diffusion: Text Guided Ground-to-Aerial View Synthesis from a Single Image using Diffusion Models

Divya Kothandaraman
dkr@umd.edu
University of Maryland College Park
USA

Tianyi Zhou
tianyi.david.zhou@gmail.com
University of Maryland College Park
USA

Ming Lin
lin@umd.edu
University of Maryland College Park
USA

Dinesh Manocha
dmanocha@umd.edu
University of Maryland College Park
USA

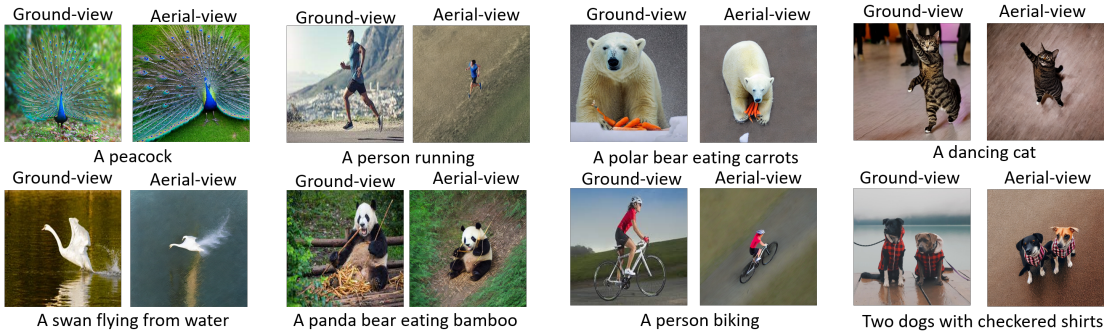

Figure 1: ***Given a single ground-view image and the corresponding text description as input,*** **Aerial Diffusion** ***generates the corresponding aerial-view image.*** **Our method does not require any supervision from aerial-view data, pairs of ground-aerial view, depth maps, semantic maps, multi-views, etc. It is one of the first approaches to achieve ground-to-aerial view translation in an unsupervised manner. We present more results and analysis in the accompanying video.**

ABSTRACT

We present a novel method, *Aerial Diffusion*, for generating aerial views from a single ground-view image using text guidance. Aerial Diffusion leverages a pretrained text-image diffusion model for prior knowledge. We address two main challenges corresponding to domain gap between the ground-view and the aerial view and the two views being far apart in the text-image embedding manifold. Our approach uses a homography inspired by inverse perspective mapping prior to finetuning the pretrained diffusion model. Aerial Diffusion uses an alternating sampling strategy to compute the optimal solution on complex high-dimensional manifold and generate a high-fidelity (w.r.t. ground view) aerial image. We demonstrate the quality and versatility of Aerial Diffusion on a plethora of images and prove the effectiveness of our method with extensive ablations and comparisons. To the best of our knowledge, Aerial Diffusion is the first approach that performs single image ground-to-aerial translation in an unsupervised manner. The full paper and code can be found at https://arxiv.org/abs/2303.11444.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

cross-view synthesis, diffusion models, text-guided, single image

ACM Reference Format:
Divya Kothandaraman, Tianyi Zhou, Ming Lin, and Dinesh Manocha. 2023. Aerial Diffusion: Text Guided Ground-to-Aerial View Synthesis from a Single Image using Diffusion Models. In *SIGGRAPH Asia 2023 Technical Communications (SA Technical Communications '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3610543.3626177

1 INTRODUCTION

The paucity of aerial data and the complexities associated with data capture from aerial cameras/ UAVs makes it difficult to train large neural networks for aerial image and video analysis. Cross-view synthesis [Regmi and Borji 2019; Tang et al. 2019] enables the synthesis of realistic aerial view images from widely available ground-view images in a controlled manner. However, cross-view synthesis requires the network to learn a very large non-trivial translation.

SA Technical Communications '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0314-0/23/12.
https://doi.org/10.1145/3610543.3626177

The network needs to hallucinate a new and enormously different view of all entities in the scene and the background, while being consistent with the details including the semantics, colors, relations between various parts of the scene, pose, etc.

Prior work [Regmi and Borji 2019; Tang et al. 2019] on ground-to-aerial generation use NeRFs and GANs. However, all of these methods use paired data for ground-view and the corresponding aerial views, which is seldom available. Moreover, training on a specific dataset leads to domain generalization issues. Instead, our goal is to develop a generic method for generating aerial views from ground-views without any paired data or other auxiliary information such as multi-views, depth, 3D mapping, etc.

While there are many diverse datasets of ground images, there are not many such good quality aerial datasets [Li et al. 2021] - hence, unpaired image-to-image translationis not a viable solution. On the contrary, text is an auxiliary modality that can be easily obtained using off-the-shelf image/video captioning tools. Moreover, text provides a natural representation space describing images. Consequently, our goal is to use the text description of a ground-view image to generate its aerial view. Recently, diffusion models have emerged as state-of-the-art architectures for text-to-image [Kawar et al. 2022; Zhang et al. 2022] high-quality realistic image synthesis. The availability of immense prior knowledge via large-scale robust pretrained text-to-image models [Rombach et al. 2022], motivates us to pose ground-to-aerial view translation as text-guided single-image translation [Kawar et al. 2022; Zhang et al. 2022]. Text-guided single-image translation methods finetune the diffusion model to the input image and then perform linear interpolation in the text embedding space to generate the desired output. However, direct application of these methods [Kawar et al. 2022; Nichol et al. 2021; Zhang et al. 2022] to ground-to-aerial translation either generates high-fidelity non-aerial images or low-fidelity aerial images.

Main contributions. We present two postulates for text-guided image translation, for ground-to-aerial translation. Based on these findings, we propose "*Aerial Diffusion*", a simple, yet effective, method for generating aerial views, given a single ground-view image and the corresponding text description as input. We apply our method on numerous in-the-wild images from various domains such as nature, animals and birds, human actions, indoor objects, etc. We conduct extensive experiments and comparisons.

2 AERIAL DIFFUSION

Notation: We use I_S and I_T to denote the ground-view and aerial-view images respectively. For the ground-view, we use the source text description txt_G = 'front view of' + txt with text embedding e_{src}. Similarly, for the aerial-view, we use the target text txt_A = 'aerial view of' + txt with text embedding e_{tgt}.

2.1 Postulates

In this section, we analyze text-guided single image translation in the context of ground-to-aerial view synthesis and present two postulates. A common strategy adopted for text-based single image translation is to use a robust text-to-image pretrained model in a two-stage process. The first step finds the 'optimized text embedding' e_{opt} (in the vicinity of e_{tgt}) that best generates the 'source' image I_S and subsequently finetune the diffusion model to generate the 'source' image I_S using e_{opt}. In the second step, a linear interpolation of e_{tgt} and e_{opt} are used to generate the edited image I_T from the finetuned neural network, i.e., the backward diffusion process is $x_{t-1} = x_t - f(x_t, t, \alpha e_{tgt} + (1-\alpha)e_{opt}),\ t = T, \cdots, 0$.

A text-based single image translation approach [Brooks et al. 2022; Hertz et al. 2022; Kawar et al. 2022; Kim et al. 2022] for ground-to-aerial generation overcomes multiple limitations in terms of data availability and generalization. However, the challenges involved in ground-to-aerial translation inhibit the direct application of existing text-based single image translation methods for ground-to-aerial generation. We present two postulates for text-based single-image translation in the context of ground-to-aerial generation.

Postulate 2.1. *Domain gap between the finetuning task (e.g., ground view generation) and target task (aerial view generation) hinders the diffusion model from generating accurate target views and introduces bias towards the source view.*

Diffusion models are probabilistic models. They are trained [Ho et al. 2020] by optimizing the negative log-likelihood of the model distribution under the expectation of the data distribution. Further simplification of the equation for formulating the training loss function involves variance reduction. In the first step of finetuning, the diffusion model is being trained to reproduce the source image given the optimized text embedding, irrespective of the input random noise. Hence, it has a natural bias towards the source image.

When the image space corresponding to the target text embedding is in the vicinity of the image space corresponding to the optimized text embedding, consistent with the variance within which the neural network was trained to generate, the generated target image is a high fidelity image consistent with the target text. When the desired transformation is large (ground-to-aerial), outside the limits of the variance, the diffusion model is unable to generate an aerial image.

Postulate 2.2. *A finetuned diffusion model cannot generalize well to the target prompt if the text embedding and image spaces corresponding to the source and the target are very different and far away from each other on the nonlinear text-image embedding manifold.*

The embedding space and the corresponding image representation space are locally linear. Hence, when the target text embedding dictates a relatively small change to the source image, a linear combination between the optimized text embedding and the target text embedding generates a high-fidelity target image, faithful to the target text. In contrast, when a linear interpolation of the text prompts is applied to ground-to-aerial translation, depending on α, the images generated are either high fidelity (but low target text faithfulness) or high target text faithfulness (but low fidelity). Moreover, the ground-view image doesn't gradually change to an oblique-view image followed by aerial-view image, the manifold is not smooth. Rather, the change is quite drastic and it is difficult to find an optimal solution in the linear interpolation space. Essentially, when there is a large perspective changes from the source to target images (i.e. involving large rotation of camera poses), the image representation space is no longer "locally linear", thereby

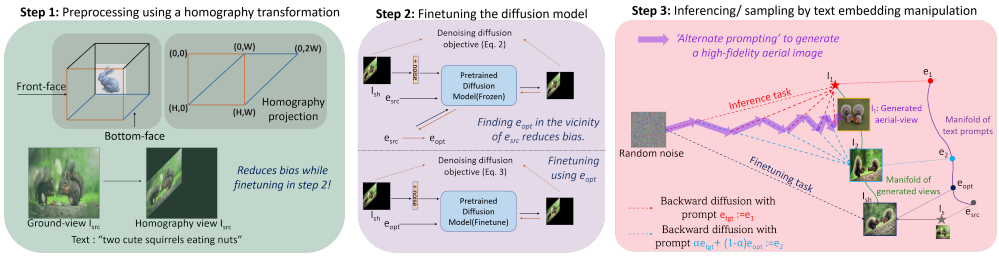

Figure 2: Aerial Diffusion.

linear interpolation is no longer adequate to generate high-fidelity images.

2.2 Method

Motivated by the challenges described above, we propose Aerial Diffusion for text guided single-image ground-to-aerial translation. An overview of our solution is as follows. We start with a pretrained robust stable diffusion [Rombach et al. 2022] model as the backbone. Our method has three stages. In the first step, we preprocess the ground-view image I_S with a carefully crafted homography transformation to generate I_{Sh}. This reduces the bias in the finetuning step. In the second step, we finetune the diffusion model by first optimizing the text-embedding within the vicinity of e_{src} to find e_{opt} that best generates I_{Sh}. Subsequently, we finetune the diffusion model to reconstruct I_{Sh}, given e_{opt}. In the third step on inferencing/sampling, we use an alternating strategy to manipulate the text embedding layer to generate a high-fidelity aerial image I_T. Next we describe each step in detail.

Step 1: Preprocessing using a homography transformation. The bias acquired by the diffusion model during the second step of finetuning inhibits large transformations. One way to decrease the bias is to reduce the number of iterations while finetuning. However, this leads to unsurprisingly low quality generated images. To decrease the bias while finetuning, we preprocess the ground-view image by transforming it with a 2D *homography transformation* [Szeliski 2022] (inverse perspective mapping). This homography projects the ground-view image to its rough 2D projected aerial view. Note that we are unable to use a 3D homography mapping to obtain the 3D aerial view projection, a better pseudo estimate of the aerial view, due to the unavailability of camera matrix, multi-views, depth information, etc. On the other hand, depth estimation methods [Fu et al. 2018; Godard et al. 2017] increase the complexity of the problem.

Consider a 3D cube (Figure 2). Without loss of generality, the 2D image captured by a ground-camera can be regarded as the projection of the scene in the front-face of the cube. A camera facing the top face of the cube will be able to capture the accurate 2D aerial view of the scene. Since we have no knowledge of the camera parameters corresponding to the ground-view image, we are unable to shift the camera to obtain a different view of the scene. With respect to the ground-camera, the 2D projection of the front-face of the cube on the bottom face of the cube is the best 'aerial projection' that we can get (inverse perspective mapping [Szeliski 2022]). This aerial projection is nowhere close to the true aerial view and does not resemble the ground-view either. Hence, when the diffusion model is finetuned, the bias is much lower than what it would have been if the optimization/finetuning were done directly with the ground-view image. This is because of the disparities between the image space of I_{Sh} and e_{src}/ e_{tgt}, ingrained in the pretrained network. Moreover, it provides a pseudo estimate of the direction in which the image needs to be transformed in order to generate its aerial view at the inference stage.

Step 2: Finetuning the diffusion model. We first optimize the text-embedding [Kawar et al. 2022; Zhang et al. 2022] to generate I_{Sh} and subsequently finetune the diffusion model using e_{opt} to generate I_{Sh}. We find e_{opt} in the vicinity of the source text embedding e_{src} - (i) the disparity between the homography transformed view and the target text is still large (though much smaller than the disparity between the ground-view and target text). Hence, it is unlikely that a good e_{opt} will be obtained when the optimization is run (around e_{tgt}) for a limited number of iterations. (ii) we do not want the network to develop a bias towards the homography image as the 'aerial view'.

To find e_{opt}, we freeze the parameters of the generative diffusion model f_θ and optimize e_{src} using the denoising diffusion objective [Ho et al. 2020]. This optimization is run for a small number of iterations, in order to remain close to e_{src} for meaningful embedding space manipulation at inferencing.To enable e_{opt} reconstruct the I_{Sh} with high fidelity, we finetune the diffusion model, again using the denoising diffusion objective [Ho et al. 2020; Saharia et al. 2022].

Step 3: Inferencing/ sampling by text embedding manipulation. Our next step is to use the finetuned diffusion model to generate a high-fidelity aerial image. Prior work [Kawar et al. 2022; Zhang et al. 2022] use linear interpolation between the optimized text embedding e_{opt} and the target text embedding e_{tgt}. Linear interpolation is not the best solution for large transformations such as ground-to-aerial generation and is unable to generate high-fidelity aerial images.

Sampling from stable diffusion [Rombach et al. 2022] involves iteratively denoising the image for T steps conditioned by text, starting with random noise. We propose to alternate between two text embeddings e_1 and e_2, starting with e_1. We designate e_1 as the target text embedding e_{tgt}. This imposes a strong constraint on the diffusion model to generate an aerial view image corresponding to the text description. The bias of the diffusion neural network motivates the network to generate an image whose details are close to the ground-view image. However, merely relying on the bias of

the neural network to capture all details of the scene is severely insufficient. Hence, we designate e_2 to be the linear interpolation of e_{opt} and e_{tgt}, controlled by the hyperparameter α. The linear interpolation can be mathematically represented as $e_2 = \alpha * e_{tgt} + (1 - \alpha) * e_{opt}$. e_2 enables the network to generate a high fidelity image while retaining the aerial viewpoint. For very low values of α, the generated image is less aerial, despite reinforcing the viewpoint to be aerial by applying e_1 alternatingly. This is because of the bias of the neural network. Very high values of α result in low fidelity images, some details of the generated aerial image are not consistent with the ground-view image. An optimal solution is by tuning α.

Linear interpolation enforces the generation of an image consistent with a text embedding in the linear space between e_{opt} and e_{tgt}. This is a reasonable when the desired change is small: when the image spaces corresponding to e_{opt} and e_{tgt} are closeby, linear interpolation works due to local linearity. When the desired change is large (such as ground-to-aerial translation), the image spaces corresponding to e_{opt} and e_{tgt} are not nearby. Since the representation spaces are not globally linear, it becomes essential to search for the solution in a much higher dimensional non-linear space. This is achieved by our alternating strategy. In summary, we manipulate the text embedding layer, such that it prioritizes fidelity and the aerial viewpoint in an alternating manner. *Alternating between text embeddings corresponding to the viewpoint and fidelity switches the denoising direction*, such that the backward diffusion takes one step towards preserving fidelity followed by another step towards generating an aerial view. As noises are gradually removed, the process ends up with a high-fidelity aerial-view image on a manifold with a better fidelity-viewpoint trade-off than linear interpolation

While the sampling repetitively alternates between e_1 and e_2, it is more beneficial to use e_1 (over e_2) at the first iteration. When the diffusion process starts with e_1, the network generates starts by generating an aerial image with details weakly dictated by its bias. On the contrary, when the diffusion process starts with e_2, the generated image in the first iteration is less aerial though with very high fidelity. The bias, along with e_{opt} serve as a strong prior towards a non-aerial viewpoint. Subsequent iterations that use e_1 are unable to overcome this strong prior to alter the viewpoint to aerial view. Hence, we start inferencing with e_1.

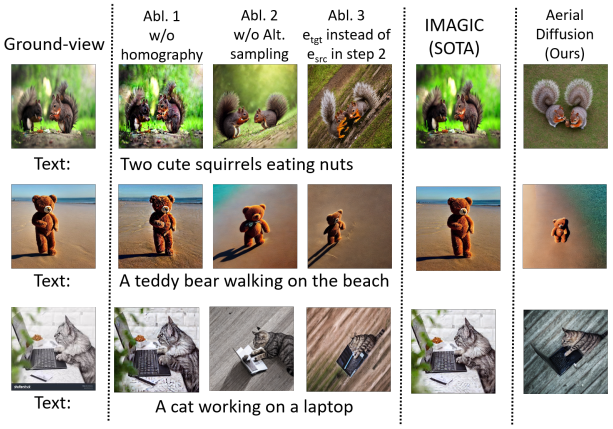

Figure 3: Ablations: we prove the effectiveness of the homography, our alternating strategy over linear interpolation and finetuning with e_{src} instead of e_{tgt}. SOTA Comparisons: IMAGIC [Kawar et al. 2022] (CVPR 2023) is unable to generate aerial views due to high bias towards the input image, domain gap and restricting the solution search to the linear interpolation manifold.

3 LIMITATIONS AND FUTURE WORK

Our method has some limitations: (i) the homography transformation results in a directional (diagonal) bias in the generated aerial image in many cases; (ii) it is limited to the knowledge contained in the pretrained stable diffusion model; (iii) the value of α needs to be manually tuned. Future work can focus overcoming these limitations. Other directions include extending Aerial Diffusion to complex scenes, generating higher-fidelity images, extending the method to videos, using the synthetic aerial data for aerial video analysis, detection, and recognition tasks.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2022. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800* (2022).

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2002–2011.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 270–279.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2022. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276* (2022).

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2426–2435.

Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. 2021. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16266–16275.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

Krishna Regmi and Ali Borji. 2019. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding* 187 (2019), 102788.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

Richard Szeliski. 2022. *Computer vision: algorithms and applications*. Springer Nature.

Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. 2019. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2417–2426.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris Metaxas, and Jian Ren. 2022. SINE: SINgle Image Editing with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2212.04489* (2022).