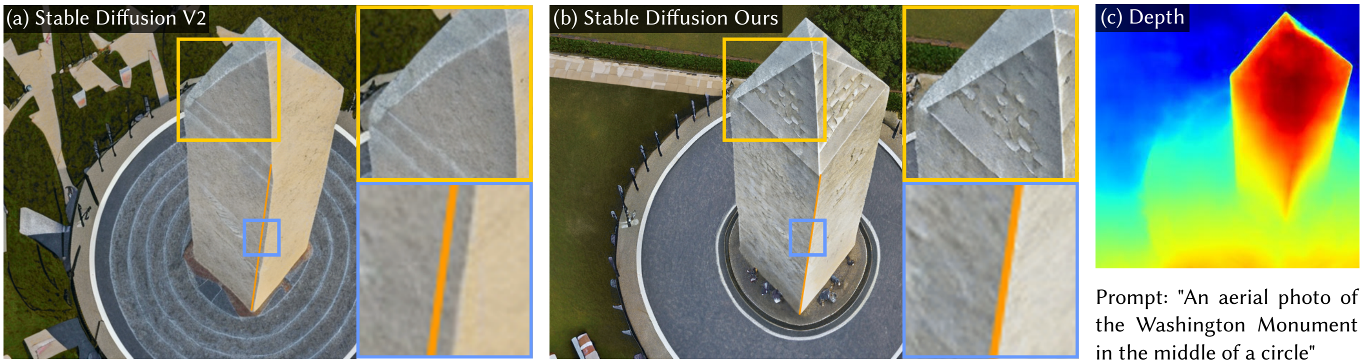
Enhancing Diffusion Models with 3D Perspective Geometry Constraints

RISHI UPADHYAY, University of California, Los Angeles, USA
HOWARD ZHANG, University of California, Los Angeles, USA
YUNHAO BA, University of California, Los Angeles, USA and Sony, USA
ETHAN YANG, University of California, Los Angeles, USA
BLAKE GELLA, University of California, Los Angeles, USA
SICHENG JIANG, University of California, Los Angeles, USA
ALEX WONG, Yale University, USA
ACHUTA KADAMBI, University of California, Los Angeles, USA



Prompt: "An aerial photo of the Washington Monument in the middle of a circle"

Fig. 1. **Images generated with our novel geometric constraint preserve straight lines and perspective.** (a) An image generated by stable diffusion v2, (b) An image generated by our fine-tuned diffusion model, and (c) the depth map and prompt both models were conditioned on.

While perspective is a well-studied topic in art, it is generally taken for granted in images. However, for the recent wave of high-quality image synthesis methods such as latent diffusion models, perspective accuracy is not an explicit requirement. Since these methods are capable of outputting a wide gamut of possible images, it is difficult for these synthesized images to adhere to the principles of linear perspective. We introduce a novel geometric constraint in the training process of generative models to enforce perspective accuracy. We show that outputs of models trained with this constraint both appear more realistic and improve performance of downstream models trained on generated images. Subjective human trials show that images generated with latent diffusion models trained with our constraint are preferred over images from the Stable Diffusion V2 model 70% of the time. SOTA monocular depth estimation models such as DPT and PixelFormer, fine-tuned on our images, outperform the original models trained on real images by up to 7.03% in RMSE and 19.3% in SqRel on the KITTI test set for zero-shot transfer.

CCS Concepts: • **Computing methodologies → Computer vision**; Neural networks.

Additional Key Words and Phrases: Diffusion Models, Perspective Constraints, Depth Estimation

ACM Reference Format:
Rishi Upadhyay, Howard Zhang, Yunhao Ba, Ethan Yang, Blake Gella, Sicheng Jiang, Alex Wong, and Achuta Kadambi. 2023. Enhancing Diffusion Models with 3D Perspective Geometry Constraints. *ACM Trans. Graph.* 1, 1 (August 2023), 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

"Re-draw The School of Athens in the style of Van Gogh", "Show an aerial viewpoint of the Washington Monument". The introduction of recent text-to-image synthesis methods such as latent diffusion models has drastically increased our creative capabilities. These models can generate anything from a Renaissance style painting to an everyday smartphone selfie from just a simple text prompt. However, as powerful as these models can be, their limited ability to adhere to physical constraints that are explicitly present in natural images restricts their potential [Wang et al. 2022]. In contrast, traditional methods of image generation such as hand-drawn art or ray-traced images place careful attention on ensuring an accurate physical environment including geometry and lighting.

Authors' addresses: Rishi Upadhyay, rishiu@ucla.edu, University of California, Los Angeles, USA; Howard Zhang, hwd15508@ucla.edu, University of California, Los Angeles, USA; Yunhao Ba, yhba@ucla.edu, University of California, Los Angeles, USA and Sony, USA; Ethan Yang, eyang657@ucla.edu, University of California, Los Angeles, USA; Blake Gella, bgella118827@ucla.edu, University of California, Los Angeles, USA; Sicheng Jiang, sicheng2020@ucla.edu, University of California, Los Angeles, USA; Alex Wong, alex.wong@yale.edu, Yale University, USA; Achuta Kadambi, achuta@ee.ucla.edu, University of California, Los Angeles, USA.

0730-0301/2023/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

One of the largest advancements in the photo-realism of hand-drawn art was the development of a system to draw accurate perspective geometry in the 1400s. While the gap between real and generated images is not as large for diffusion models as it was back then, a greater consideration for perspective accuracy can have a similarly large impact in the photo-realism of their outputs.

Perspective is one of the most important physical constraints because it ensures object properties such as size, relative location, and depth are accurately represented. In a sense, it ensures physical accuracy [Kadambi 2020]. This allows the use of perspective accurate data for downstream tasks such as camera calibration [Beardsley and Murray 1992; Caprile and Torre 1990; Chen and Jiang 1991; He and Li 2007; Li et al. 2010], 3D reconstruction [Guillou et al. 2000; Wang et al. 2009], scene understanding [Geiger et al. 2014; Han and Zhu 2009; Satkin et al. 2012], and SLAM [Camposeco and Pollefeys 2015; Georgis et al. 2022; Lim et al. 2022].

However, current diffusion based image generators such as [Bau et al. 2021; Radford et al. 2021; Razavi et al. 2019; Rombach et al. 2022b; Yu et al. 2022] do not generate perspectively accurate data [Farid 2022b]. Please refer to Fig. 1(a) for an example of this phenomenon. This is because latent diffusion models typically lack the interpretability necessary for explicit encoding of a physical prior such as perspective in the model architecture [Kadambi et al. 2023]. By utilizing a novel loss function that ensures the gradient field of an image aligns with its expected vanishing points, we are able to encode this physical prior. By enforcing this perspective prior on generated images, we also increase the accuracy of object properties important for downstream computer vision tasks and photo-realism.

As it turns out, the perspective correctness of an image has a strong influence over its overall scene coherence and therefore realism. This is most likely true because, as mentioned before, perspective provides crucial information regarding the size, relative location, and depth of a scene. To illustrate this, we set up a human subjective test where the photo-realism of our perspective-corrected images is put to the test. We show that latent diffusion models which utilize our novel perspective loss generate images that are rated as more realistic an overwhelming majority of the time as compared to images generated by the base diffusion model. We also verify the visual benefits of our proposed constraint by applying it to the inpainting task. We show that inpainted images generated from models trained with our loss consistently appear more perceptually similar to the original image than images from models without our loss.

Additionally, images generated with our novel loss prove beneficial to the accuracy of downstream tasks which are inherently reliant on these same object properties. As proof of this concept, we fine-tune multiple SOTA monocular depth estimation models such as DPT [Ranftl et al. 2021] and PixelFormer [Agarwal and Arora 2023]. We show that training on data with accurate perspective leads to models with higher performance that can capture high-frequency details to a higher degree.

1.1 Contributions

In summary, we make the following contributions:

- We introduce a novel geometric constraint on the training process of latent diffusion models to enforce perspective accuracy.
- We show that images from models trained with this constraint appear more realistic than models trained without this constraint 69.6% of the time.
- We demonstrate that downstream tasks which benefit from more geometrically accurate inputs (such as monocular depth estimation) improve by up to 7.03% in RMSE and 19.3% in SqRel.

2 RELATED WORK

2.1 Synthetic Image Generation

Image generation, while a popular task, has proven to be difficult because of the high dimensional space and variety of images. One of the most popular techniques for image generation has been Generative Adversarial Networks (GANs) [Goodfellow et al. 2020]. While GANs are capable of high quality image synthesis [Brock et al. 2019], they are limited by the fact that they are difficult to train, often failing to converge or collapsing into a mode where all generated images are the same [Arjovsky et al. 2017; Mescheder et al. 2018]. Another popular image generation technique is Variational Auto-encoders (VAEs) [Kingma and Welling 2014] which have stronger theoretical guarantees, but cannot match GANs in image quality [Child 2021; Vahdat and Kautz 2020]. Recently, diffusion models [Sohl-Dickstein et al. 2015] for image generation have grown in popularity. These models work by reversing a diffusion process which adds noise to high quality images and are capable of generating high quality samples from a variety of distributions [Daras et al. 2022; Dhariwal and Nichol 2021; Ho et al. 2020]. Subsequent works have expanded the scope even further by adding text guidance to the diffusion process [Ramesh et al. 2022; Saharia et al. 2022], simplifying the inverse process [Wallace et al. 2022], and reformulating the diffusion process to occur in a latent space for speed benefits [Rombach et al. 2022b]. While recent work has explored guiding diffusion models in various ways [Ho and Salimans 2022; Meng et al. 2023; Rombach et al. 2022a; Wallace et al. 2023], most diffusion models rely almost entirely on their vast datasets and text encoders for priors on scene composition and object properties. Some , but this work generally focuses on the whether or not objects are present as opposed to scene physics. This means that there are no explicit guarantees that generated images will be physically accurate, making them a poor fit for use in synthetic datasets. Our work aims to add 3D geometry constraints to image generators in order to improve the quality of generated images.

2.2 Vanishing Points in Computer Vision

Vanishing points have many varied and important uses in computer vision. One common use for vanishing points is camera calibration. Early examples of this include [Beardsley and Murray 1992; Caprile and Torre 1990; Chen and Jiang 1991] who use vanishing point geometry to compute the intrinsics and extrinsics of one or more cameras given single or multiple images. Subsequent papers, such as [He and Li 2007; Li et al. 2010], provided improved techniques that were simpler or required less data and assumptions. In addition, newer works

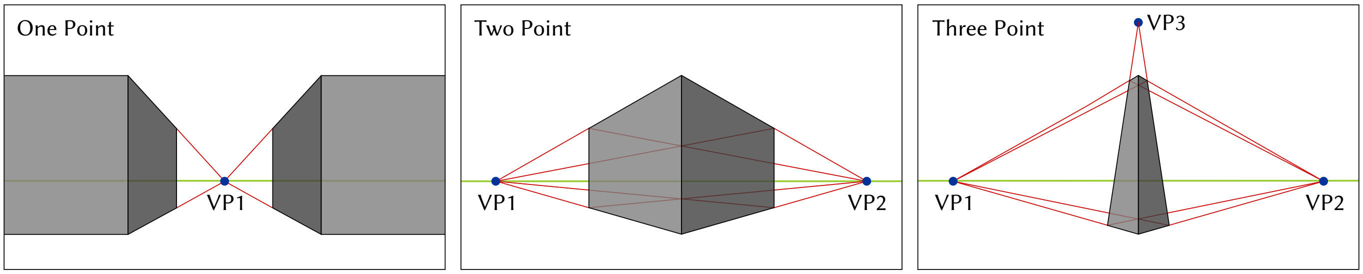


Fig. 2. **Examples of one, two, and three-point linear perspective.** Vanishing points are labeled in blue, perspective lines are in red, and the horizon lines are in light green. One-point perspective is typically used when there is one focal point of the image or when only one side of an object is visible. Two-point perspective is used to illustrate multiple sides of an object, while three-point perspective is used for viewpoints that are above or below the horizon line of the 3D scene.

began to not only compute camera parameters, but also use them to compute 3D reconstructions of single images [Guillou et al. 2000; Wang et al. 2009]. Beyond camera calibration, vanishing points are also useful for general scene understanding. [Han and Zhu 2009] use vanishing points to help create generative grammar for synthetic scenes, [Geiger et al. 2014] use vanishing points as priors for 3D scene and traffic understanding, and [Satkin et al. 2012] estimate 3D models from singular images using vanishing point priors. Vanishing points are also particularly useful for road detection thanks to easily identifiable perspective lines, as demonstrated by [Kong et al. 2009; Liou and Jain 1987]. Vanishing points are also regularly used in SLAM techniques. [Lee et al. 2009] were one of the first in this space, using vanishing points to identify the heading of a robot for navigation. Subsequent works further expanded the capabilities of SLAM systems built on vanishing points including [Camposeco and Pollefeys 2015; Georgis et al. 2022; Lim et al. 2022] who use vanishing points to identify direction and perform structural mapping of scenes in real-time. Given the significance of vanishing points in computer vision, we aim to enhance image generators with accurate perspective, in order to benefit photo-realism and downstream tasks.

2.3 Monocular Depth Estimation

Supervised methods for monocular depth estimation typically require paired image and depth data. One of the first works in this area was Make3D [Saxena et al. 2008] which relied on hand-crafted features and Markov random fields. Subsequent works then applied deep learning to the problem, starting with multi-scale convolutional networks [Eigen et al. 2014] and followed by conditional random fields [Li et al. 2015], residual networks [Laina et al. 2016], convolutional neural fields [Liu et al. 2015; Xu et al. 2018], and most recently transformers [Agarwal and Arora 2023; Ranftl et al. 2021, 2020]. Many approaches also take advantage of known geometric relationships, such as normals [Qi et al. 2018] and planes [Lee et al. 2019; Yang and Zhou 2018]. Newer techniques have also taken an unsupervised approach [Fei et al. 2019; Wong and Soatto 2019] or use multi-modal data capture [Singh et al. 2023]. However, most supervised monocular depth estimation models are limited by the availability of paired data on which to train as this data is difficult to collect.

In order to overcome the challenge of a lack of sufficient training data, many techniques turn to synthetic datasets. The renderers used to generate the images in these datasets can often generate corresponding ground-truth data, making it simple to acquire pixel-aligned ground-truth depth maps. In addition, these renderers often allow for different types of data, such as varied weather conditions or indoor vs. outdoor scenes, making them an attractive way to get training data. Examples of such datasets include Virtual KITTI, a photorealistic copy of the popular self-driving dataset KITTI [Gaidon et al. 2016; Geiger et al. 2013] and SYNTHIA, a dataset that includes depth and semantic segmentation information for images of a synthetic city [Ros et al. 2016; Zolfaghari Bengar et al. 2019]. Although these datasets are often quite realistic, there are often key differences between synthetic and real images which leads models trained on synthetic images to achieve lower performance when tested on real datasets compared to models trained and tested on real images. This difference in performance is referred to as the Sim2Real gap. As monocular depth estimation is a popular task, many works have attempted to address the problem of the Sim2Real gap [Cao et al. 2018; Damodaran et al. 2018; Long et al. 2015; Rozantsev et al. 2018; Sankaranarayanan et al. 2018]. However, all of these techniques approach the problem by attempting to improve the neural network architectures. On the other hand, we approach this problem from the perspective of improving the synthetic data used to train the neural networks.

In addition to monocular depth estimation, the techniques we describe in the paper can be easily applied to the task of depth completion as well since the data format is the same for both tasks [Nazir et al. 2022; Wong et al. 2021; Wong and Soatto 2021].

3 PERSPECTIVE BACKGROUND

3.1 Linear Perspective

Although perspective is a word commonly used in a variety of contexts, it has a very specific meaning in terms of art and photography: techniques used to draw objects in 2D such that their 3D attributes are correctly modeled. In practice, perspective refers to a multitude of different techniques which can be used to create a 3D feel, but the most common technique is called linear perspective. In linear perspective, all mutually parallel lines, on the same or parallel planes, in 3D space, converge to a single point in the image plane which

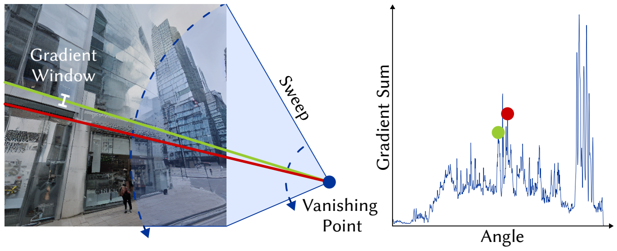


Fig. 3. **Graphical description of our geometric constraint.** *Left:* A visualization of how the loss function sweeps lines across the image. *Right:* $D(v, \mathbf{x})$ plotted for the image at right. The red and yellow lines in the left plot are identified by the corresponding dots.

is referred to as a vanishing point. The only exception to this rule is sets of lines that are exactly parallel to the camera sensor. In this case, these lines are also parallel in the image plane. A typical drawing/image often has anywhere from one to three vanishing points, with the number of vanishing points determining the style and view of the drawing/image. Another key component of linear perspective is the horizon line. The horizon line is a horizontal line that represents the viewer's eye level in an image, and typically at least one of the vanishing points of an image lies on this line. A visualization of these principles can be found in Fig. 2.

3.2 Perspective Consistency in Images

Perspective in images is not always easy to confirm, as the vanishing points of an image can only be easily identified with the aid of parallel lines in 3D space, which may not always exist in images. For images that do have sets of parallel lines, perspective consistency can be verified by extending sets of parallel lines in either direction until they intersect and ensuring that all pairs of lines in a set intersect at the same point.

Natural Images. By the math of perspective projection for a pinhole camera, an point $\mathbf{X} = (X, Y, Z)$ is projected to a point $\mathbf{x} = (x, y) = (fX/Z, fY/Z)$ [Ma et al. 2003]. If we are concerned with a line $L = \mathbf{O} + t\mathbf{D}$, after replacing X, Y, Z above with the equation for a line and taking the limit as t goes to positive/negative infinity, we see that the final projected point is dependent on only $\mathbf{D}$. Therefore, sets of parallel lines, that are not parallel to the camera plane, will all come together to the same point, known as a vanishing point.

Synthetic Images. Although natural images are forced to follow perspective rules, there are no such restrictions on synthetic images, particularly images generated by deep learning approaches. Most of the loss functions used to train these models are focused on image quality or how well prompts are followed, meaning physical properties such as perspective, shadows, or lighting can often be neglected [Farid 2022a,b]. An example of this can be seen in Fig. 1(a).

4 IMPROVING PERSPECTIVE ACCURACY OF GENERATED IMAGES

Our fine-tuned model is built on top of the latent diffusion models introduced by [Rombach et al. 2022b], using code from [Pinkney 2022]. We describe the latent diffusion process in Section 4.1. We add a new term to the traditional loss function and train on a specialized dataset that provides ground truth vanishing points. This new constraint is described in Section 3.2.

4.1 Latent Diffusion Models

Traditional image generation diffusion models are concerned with a forward diffusion process over images $\mathbf{x}_0,...,\mathbf{x}_T$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where q is the forward diffusion function, t is the current time step, and $\mathbf{I}$ is the identity. $\alpha_t = 1 - \beta_t$ and $\beta_1,...,\beta_T$ compose a pre-selected variance schedule. The reverse process is then parameterized as:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma(\mathbf{x}_t, t)), \tag{2}$$

where p is defined as the reverse diffusion function and $\Sigma(\mathbf{x}_t, t)$ is typically set to time-dependent constants. $\mu_\theta(\mathbf{x}_t, t)$ is defined as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right), \tag{3}$$

where $\overline{\alpha}_t = \Pi_{i=1}^t \alpha_i$, and $\epsilon_\theta(\mathbf{x}_t, t)$ is a learned function parameterized by a UNet model [Ronneberger et al. 2015] with learned parameters θ. Based on this, the traditional diffusion model loss is as follows:

$$L_{\text{DM}} = \mathbb{E}_{\mathbf{x},\epsilon\sim\mathbb{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right]. \tag{4}$$

More details and derivations can be found in [Ho et al. 2020]. Latent diffusion models work very similarly, but perform the forward and reverse diffusion processes in latent spaces. Specifically, an encoder and decoder are introduced to translate to and from the latent space. The encoder is defined as: $\mathcal{E} : X \in R^{H\times W\times 3} \mapsto Z \in R^{h\times w\times 3}$, while the decoder is defined as: $\mathcal{D} : Z \in R^{h\times w\times 3} \mapsto X \in R^{H\times W\times 3}$, where $h = H/f$, $w = W/f$ and f is a downsampling factor. With this formulation, the loss function now becomes:

$$L_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(\mathbf{x}),\epsilon\sim\mathbb{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right], \tag{5}$$

where the image x_t is replaced by its latent space representation z_t.

In order to add perspective priors to a latent diffusion model, we add an additional perspective loss term. At a high level, this loss works by sweeping lines extending out from a vanishing point over the image and calculating the sum of image gradients across the line, as illustrated in Fig. 3. Pseudocode for this algorithm is shown in Alg. 1. This sum is designed to represent how "edge-like" the region along that line is in the image. We can then write our new loss as:

$$\begin{aligned} L_{\text{DM}} = \mathbb{E}_{\mathcal{E}(\mathbf{x}),\epsilon\sim\mathbb{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right] + \\ \lambda\mathbb{E}_{\mathbf{x},\epsilon\sim\mathbb{N}(0,1),v}\left[L_{\text{persp}}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{v}_\mathbf{x})\right]. \end{aligned} \tag{6}$$

where λ is a weight factor for our perspective loss, $\mathbf{v}_\mathbf{x}$ is a set of vanishing points in image space and $\hat{\mathbf{x}}$ is our reconstructed image, which can be written as:

$$\hat{\mathbf{x}} = \mathcal{D}\left(\frac{1}{\sqrt{\overline{\alpha}_t}}\left(\mathbf{z}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta(z_t, t)\right)\right). \tag{7}$$

where t is randomly chosen between 0 and T for each iteration. In order to define L_{persp}, we first define some intermediate quantities:

- $G_\mathbf{x}$ represents the gradients of an image $\mathbf{x}$ computed with a 3x3 Sobel filter.

ALGORITHM 1: Algorithm to compute perspective loss

```
Function perspective_loss($\mathbf{x}, \hat{\mathbf{x}}, \mathbf{v_x}$)
    Input  : Image $\hat{\mathbf{x}}$
    Input  : Ground Truth image $\mathbf{x}$
    Input  : Vanishing Points $\mathbf{v_x}$
    $G_{\mathbf{x}} \leftarrow$ img_derivative($\mathbf{x}$)
    $G_{\hat{\mathbf{x}}} \leftarrow$ img_derivative($\hat{\mathbf{x}}$)
    $loss \leftarrow 0.0$
    foreach $v \in \mathbf{v_x}$ do
        $\phi_{\min}, \phi_{\max} \leftarrow$ calc_image_angle($v$)
        for $i \leftarrow 0; i < N; i = i + 1$ do
            $angle \leftarrow \frac{i}{N} * (\phi_{\max} - \phi_{\min}) + \phi_{\min}$
            $d \leftarrow$ calc_perp_vec($angle$)
            $p \leftarrow$ get_line_pixels($v$, $angle$)
            $D(i) \leftarrow \sum_p |G_{\hat{\mathbf{x}}} \cdot d|$
            $D_{\text{gt}}(i) \leftarrow \sum_p |G_{\mathbf{x}} \cdot d|$
            $loss \leftarrow loss + \text{norm}(D - D_{\text{gt}})$
        end
        $loss \leftarrow loss/|\mathbf{v}|$
    end
    return $loss$
end
```

- $\phi_{\min}$ and $\phi_{\max}$ represent the minimum and maximum angle from the vanishing point to a corner of the image relative to the x-axis.
- ϕ_0,...,ϕ_n represent n equally-spaced angles between $\phi_{\min}$ and $\phi_{\max}$.
- v represents a particular vanishing point in the set $\mathbf{v_x}$.
- $l_i(v, k)$ represents a point at time k on a ray $l_i(v)$ starting at v in the direction of ϕ_i.
- $d_i(v)$ represents a vector perpendicular to the line $l_i(v)$.

Using these, we define:

$$D_i(v, \mathbf{x}) = \int_{k_0}^{k_1} d_i \cdot G_{\mathbf{x}}(l_i(v, k))dk, \tag{8}$$

where k_0 and k_1 represent the times of the intersection of $l_i(v)$ with $\mathbf{x}$. $D_i(v, \mathbf{x})$ is then our measure of how "edge-like" the region along this ray is, and we can then define:

$$L_{\text{persp}}(\hat{\mathbf{x}}, \mathbf{x}, \mathbf{v_x}) = \frac{1}{|\mathbf{v_x}|} \sum_{v \in \mathbf{v_x}} ||D(v, \hat{\mathbf{x}}) - D(v, \mathbf{x})||_2. \tag{9}$$

In practice, the integral in Eq. 8 becomes a sum over the image pixels that the line intersects.

5 EXPERIMENTS

In order to evaluate our proposed constraint, we conduct comprehensive experiments. In Section 5.1, we detail how we fine-tune latent diffusion models with the proposed constraint, in Section 5.2, we detail how we fine-tune monocular depth estimation models on images generated from our fine-tuned models. In Section 5.3, we describe how we evaluate the photo-realism of images generated from our fine-tuned models, and in Section 5.4, we describe our ablation studies.

5.1 Training Latent Diffusion Models

For all of our image generation experiments, we build off the depth-conditioned Stable Diffusion V2 model from [Rombach et al. 2022b]. This model is trained on LAION 5B, a database of 5.85 billion image caption pairs [Schuhmann et al. 2022]. In this paper, we refer to this model as the baseline model.

Datatsets. In order to fine-tune the baseline model, we use the HoliCity dataset [Zhou et al. 2020]. This dataset provides 50,078 real images taken in London along with ground truth vanishing points for each image. We use MiDaS [Ranftl et al. 2020] to compute a depth prediction for each image which is then used as conditioning for the latent diffusion model.[1] This is the same procedure used to originally train the depth-conditioned model [Rombach et al. 2022b]. Captions used for conditioning are generated for each image using the BLIP captioning model [Li et al. 2022].

Training Details. The code for our fine-tuned model is built using PyTorch on top of [Pinkney 2022], which is built on top of the original code released by [Rombach et al. 2022b]. The original code from [Pinkney 2022] is built on top of Stable Diffusion v1, so part of the modifications made by us include updating the code to be compatible with Stable Diffusion v2 checkpoints, including updating the encoder/decoder and dataloaders. We update the loss function of the baseline model to the loss function detailed in Eq. 6. We train at an image resolution of 512×512 with a learning rate of 1e-6 and $\lambda = 0.01$. We train for 4 epochs or approximately 200k steps with an effective batch size of 16 after gradient accumulation. At this point, the perspective loss had saturated. All models were trained on 4 RTX3090 GPUs. Results are shown in Section 6.1.

5.1.1 *Inpainting.*

In addition to text-to-image generation, we also test the value of our constraint for the inpainting task where a model is asked to fill in masked regions of an image. Applying our proposed constraint to the inpainting task does not require any extra training, as we are able to take our general text-to-image diffusion models and perform inpainting using the techniques described by [Lugmayr et al. 2022]. We evaluate the results using the LPIPS metric [Zhang et al. 2018] as is the norm for the inpainting task. LPIPS measures the perceptual similarity between two images using features from deep neural networks, in particular AlexNet. Results are shown in Fig. 7 and Table 4 and are discussed in Section 6.1.1

5.2 Training Monocular Depth Estimation Models

In order to evaluate the performance from another perspective, we also conduct an experiment on the effect of our new images on monocular depth estimation models. In particular, we fine-tune DPT-Hybrid [Ranftl et al. 2021] and PixelFormer [Agarwal and Arora 2023] on images generated from both the baseline model and our fine-tuned model. DPT-Hybrid is originally trained on MIX 6, a collection of 10 datasets described in [Ranftl et al. 2021], and PixelFormer is originally trained on the KITTI dataset. In order to generate these images, we rely on the SYNTHIA-AL [Zolfaghari Bengar et al. 2019] and Virtual KITTI 2 [Cabon et al. 2020; Gaidon et al.

[1] The HoliCity dataset also provides ground truth depth images, however, they are derived from a CAD model, meaning they are missing finer details such as people, cars, and trees.

2016] datasets. SYNTHIA-AL contains 70,000 images and Virtual KITTI 2 contains 2,656 images. We take only depth maps from both datasets, and use them as conditioning to generate synthetic images using the base, and our latent diffusion models. In addition, we use BLIP [Li et al. 2022] to generate captions for all images. For Virtual KITTI 2, we take 8 random crops per image. We also generate diffusion images with 4 different seeds, resulting in a total of 84,992 images derived from the Virtual KITTI 2 dataset. For SYNTHIA, we use the original images, resulting in a total of 70,000 images. Combined, our dataset is 154,992 images and covers various city and driving scenes. We refer to this dataset as All. We additionally train the depth estimation models on images generated only from vKITTI, and refer to this dataset as vKITTI. We additionally append the name of the model used to generate different datasets so that All Enhanced refers to the full set of 155k images generated by our Enhanced model while All Base refers to the full set of images generated by the Baseline model. Results of fine-tuning on these datasets are discussed in Section 6.2.

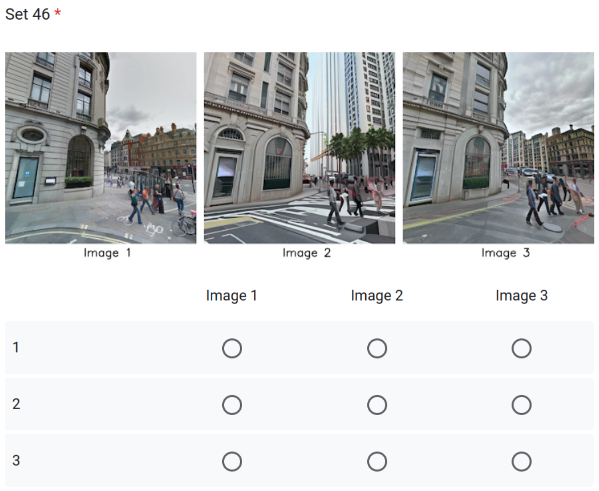
Training Details. For DPT-Hybrid, we train with a learning rate of 5e-6 for 19,500 steps with a batch size of 16. We use a scale and shift invariant loss as described in [Eigen et al. 2014; Ranftl et al. 2021]. For PixelFormer, we train with a learning rate of 4e-6 for 20,800 steps with a batch size of 8. We train on 1 RTX3090 GPU using the same loss as DPT.

Test Sets. We evaluate the trained depth estimation models on commonly used real datasets KITTI [Geiger et al. 2012] and the outdoor subset of DIODE [Vasiljevic et al. 2019]. We use the Eigen split for KITTI [Eigen et al. 2014] and a test set of 500 images from DIODE.

Metrics. In order to evaluate the performance of the models, we follow the procedure used by [Ranftl et al. 2021] and we adopt common depth estimation metrics: Absolute relative error (Abs Rel), Square relative error (Sq Rel), Root mean squared error (RMSE), Log RMSE (RMSE log), and Threshold Accuracy (δ_i) at thresholds τ_i's = 1.25, 1.25^2, 1.25^3 as used in [Agarwal and Arora 2023; Ranftl et al. 2021, 2020].

5.3 Human Subjective Test Methodology

In order to evaluate the photo-realism of images generated by our fine-tuned models, we run human subjective tests on the Prolific [Academic Ltd 2023] website. We ran two tests, one comparing our enhanced model with the baseline model and one comparing our enhanced model with an ablation model. We set up the test as a ranking task where participants are asked to rank sets of three images (Real, Baseline, Ours or Real, Ablation, Ours) in order of photo-realism. The real images come from the HoliCity dataset [Zhou et al. 2020], a landscapes dataset from Kaggle [Rougetet 2020], and an animal images dataset from Kaggle [Awais 2020]. The baseline, ablation, and enhanced (ours) images are generated using depth maps extracted from the real image by MiDaS [Ranftl et al. 2020] and prompts from the BLIP captioning model [Li et al. 2022]. Participants were shown all three images side by side in random order. Please refer to Fig. 4 for a visualization of the testing setup. We recruit 50 participants across the world and ask them to rate 80 sets of images. Participants were given up to 90 minutes to complete the task. Results from this test are in Section 6.3 and Fig. 11.



Set 46 *

	Image 1	Image 2	Image 3
1	○	○	○
2	○	○	○
3	○	○	○

Fig. 4. **A screenshot of the graphical user interface for the human subjective test we performed on the Prolific platform.** Annotators are asked to rank the image by realism, with "1" being the most and "3" being the least real. Images include one generated from a baseline model, one generated from our enhanced model, and one real image in random order.

5.4 Ablation Study

In order to evaluate the benefits of our proposed constraint, we additionally fine-tune the baseline model on the same dataset but without our updated loss. We refer to this model as the Ablation model. We additionally generate the same synthetic datasets and train the same monocular depth estimation models described in Section 5.2. Results are shown in Section 6.4. An ablation study was also done for the human subjective tests and for the inpainting task. Results are described in Section 6.3 and shown in Fig. 11, Fig. 7, and Table 4.

6 RESULTS

This results section is split into sub-sections according to the experiments described in Section 5. In Section 6.1, we describe the results of fine-tuning latent diffusion models. In Section 6.2, we discuss the results of fine-tuning SOTA monocular depth estimation models on our generated images. In Section 6.3, we discuss the results of our human subjective test, and in Section 6.4, we discuss the results of our ablation tests.

6.1 Fine-tuned Latent Diffusion Models

We show some representative generations from our fine-tuned model in Fig. 5. In the figure, we show the depth maps used to condition the diffusion models along with generations from the baseline model and our enhanced model. Images from the baseline model tend to suffer from curved lines and distortions that affect perspective accuracy. In particular, the baseline model tends

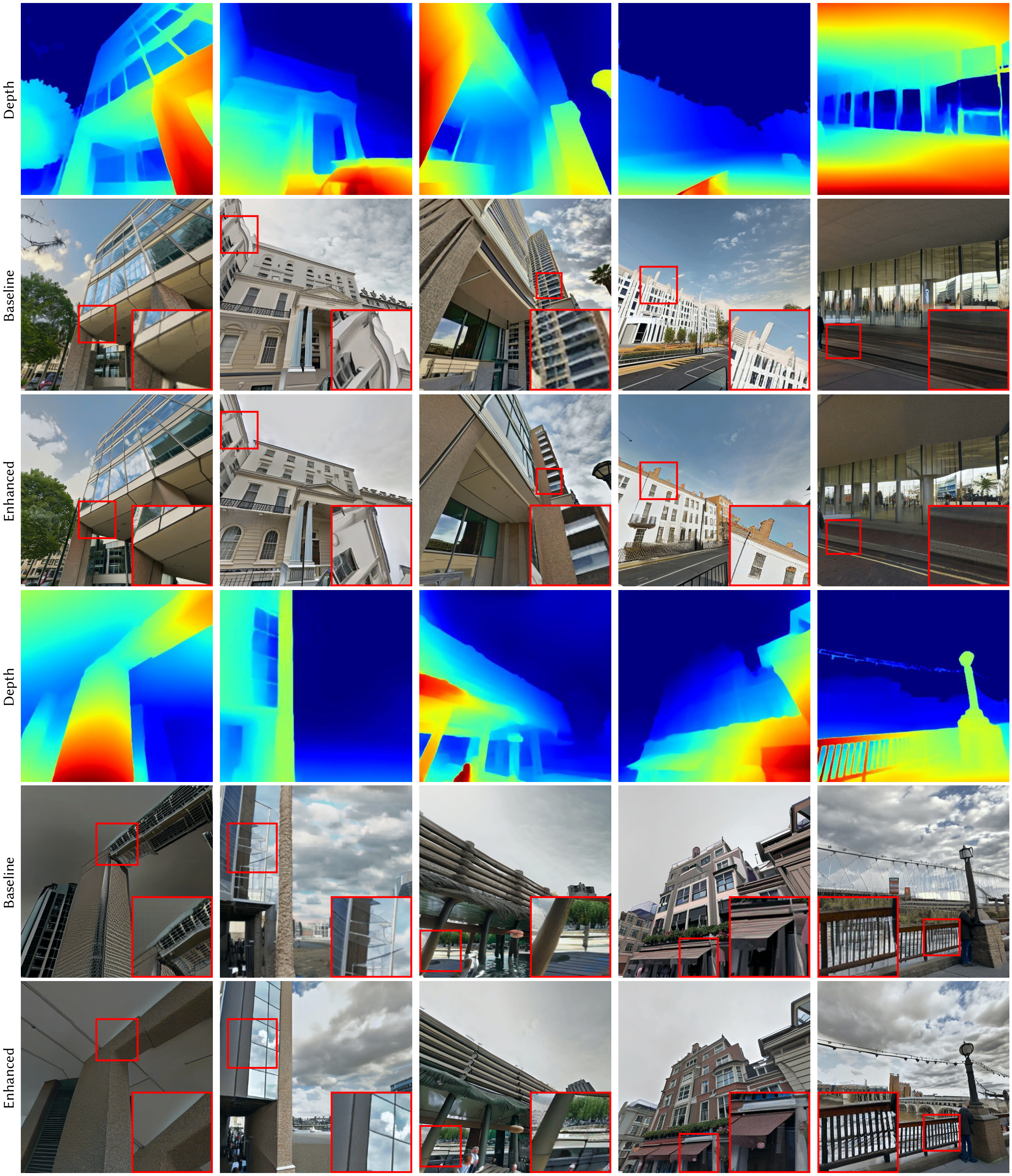


Fig. 5. **Images from our model are better at preserving straight lines.** Examples of outputs from the base model and from our enhanced model. The depth maps these outputs are conditioned on are put at the top. Inlets show specific regions of interest.

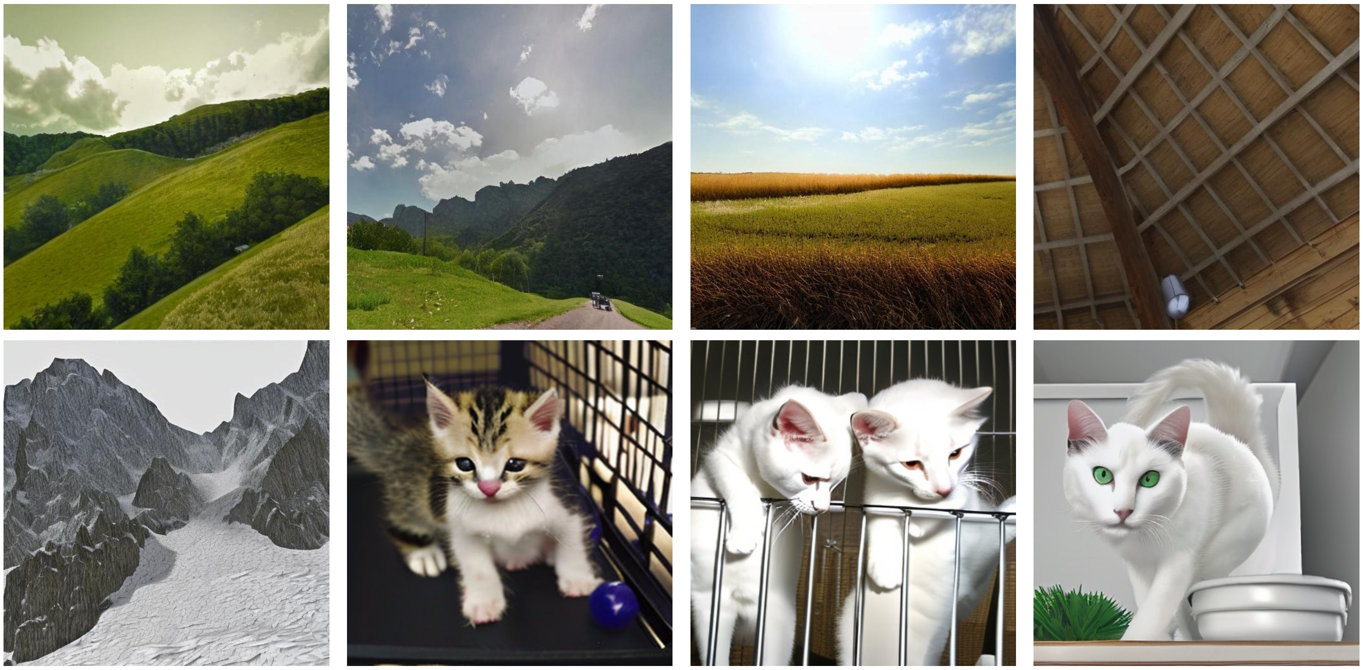


Fig. 6. **Despite being fine-tuned on images of city scenes, our model is able to generate high-quality images of varied settings including nature landscapes, indoor scenes, and pictures of animals.** Images were taken from a landscapes dataset [Rougetet 2020], an animal dataset [Awais 2020], and the indoor subset of the DIODE dataset [Vasiljevic et al. 2019].

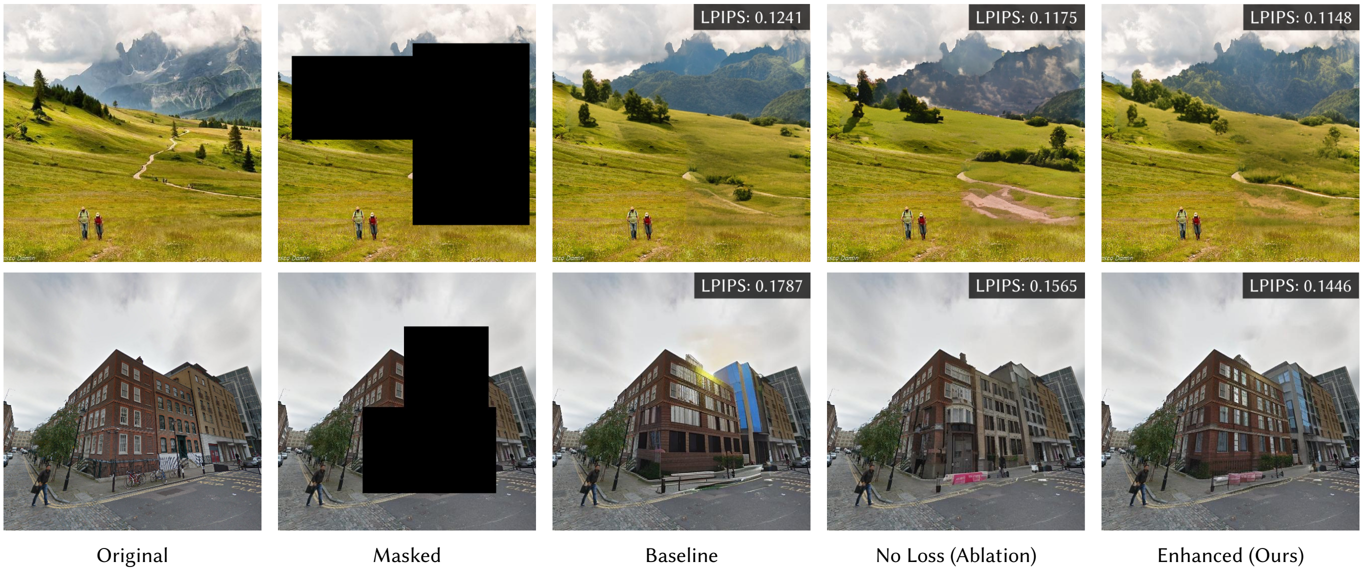


Fig. 7. **The proposed geometric constraint provides benefits for the inpainting task on diverse scenes.** Images reconstructed with our enhanced model consistently outperform the baseline and ablation models on LPIPS scores (shown in the top right, lower is better).

to have trouble accurately generating regions with windows, high-frequency details such as many parallel horizontal or vertical lines, and corners. We also draw perspective lines on images from the baseline and our models in Fig. 8. Images from our model tend to have more coherent perspective lines and more accurate vanishing points. In addition, in both figures, because of the aforementioned distortions, the baseline images look further from the distribution of natural images than images from our model. Since our enhanced model is fine-tuned on a dataset of mainly only cityscapes, we also generate varied nature, animal, and indoor scenes to verify that this fine-tuning does not limit the ability of the model to generate other types of images. Some representative images are shown in Fig. 6.

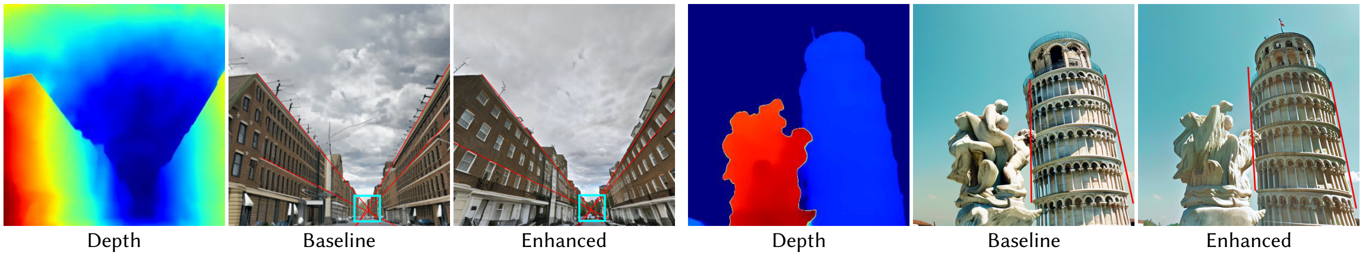


Fig. 8. **Images from our model have more consistent vanishing point lines.** This figure shows examples of stable diffusion outputs from the baseline model and from our model with perspective loss along with perspective lines for the image. The depth maps these outputs are conditioned on are put in the left-hand column. Note that for the baseline image in the first row, the lines do not intersect at a single vanishing point, violating perspective geometry. These violations can sometimes result in curved lines as seen in the baseline image in the second row.

6.1.1 Inpainting. We evaluate the inpainting performance of our models using both qualitative (Fig. 7) and quantitative (Table 4) results. All three models of interest, the baseline model, ablation model, and enhanced model were tested on the combination of two datasets: the HoliCity validation set [Zhou et al. 2020] and a landscape dataset [Rougetet 2020]. The LPIPS metric [Zhang et al. 2018], which measures perceptual similarity using features from deep image networks, was used to compare models as is the norm for the inpainting task. We used the official implementation provided by [Zhang et al. 2018]. Note that lower is better for the LPIPS metric. As seen in Table 4, our enhanced model consistently outperforms both the baseline model and ablation model, with a 7.1% improvement over the baseline model and a 3.6% improvement over the ablation model on the combined dataset.

6.2 Monocular Depth Estimation

In order to evaluate the performance of our fine-tuned depth estimation models, we use both qualitative and quantitative measures. A qualitative comparison is shown in Fig. 9, while quantitative comparisons are in Table 1 and Table 2.

DPT-Hybrid. We fine-tune one model from the base DPT-Hybrid using the generated vKITTI datasets and then test the model on both the original KITTI test set (Eigen Split) and a subset of the DIODE Outdoor test set. Results are in Table 1. The models fine-tuned on images generated from our diffusion model outperform the original DPT-Hybrid model on all metrics on both datasets and outperform the model fine-tuned on images generated by the baseline model on all metrics for KITTI and all but one metric (SqRel) for DIODE Outdoor. In addition, for the DIODE Outdoor dataset, the original DPT-Hybrid model outperforms the base model on five out of eight metrics, but outperforms our model on no metrics. In particular, our model shows a 7.03% improvement in RMSE and a 19.3% improvement in SqRel over the original model while also demonstrating a 3.4% improvement in SqRel and a 2.2% improvement in SiLog over the baseline model. Fig. 9 also shows qualitative comparisons between the original DPT-Hybrid model and the model fine-tuned on images generated by our enhanced diffusion model. Each set of images contains the input image, ground truth depth map (dilated with a 3×3 kernel), and error maps from both the original model and our enhanced model. Additionally, the RMSE values for each of the depth predictions are shown in the top right of the error maps. The depth models from our model capture more high-frequency detail such as corners and poles, and also consistently have lower RMSE values.

PixelFormer. We fine-tune the base PixelFormer using both the generated vKITTI dataset and the full generated dataset and evaluate on the DIODE Outdoor test set. Results are shown in Table 2. The model fine-tuned on images from our diffusion model outperforms the original model and the models trained on images from the baseline model on all metrics. Our model trained on the vKITTI dataset achieves a 4.1% improvement in RMSE over the original model, while our model trained on the entire dataset achieves an 11.6% improvement in SiLog over the original model and a 2.4% improvement over the model trained on baseline images. Additionally, the original model outperforms the baseline model trained on the entire dataset on five of eight metrics, but outperforms the model trained on our images on no metrics.

6.3 Human Subjective Tests

Results from the human subjective tests are shown in Fig. 11. (a) shows the comparison between our enhanced model and the baseline model while (b) compares our enhanced model and the ablation model. Over all trials, images from our enhanced model appear more photo-realistic than images from the baseline model 69.6% of the time and appear more photo-realistic than images from the ablation model 67.5% of the time. In addition, the average rank of our images (between 1 and 3, lower is better) compared to the baseline was 1.9345 vs 2.4383 and was 1.9584 vs 2.4011 compared to the ablation model. The differences in average rank between our enhanced images and the baseline images (0.5038) and the difference between our images and the ablation images (0.4427) are also consistently less than the difference in average rank between our enhanced images and real images (0.3072 and 0.318 respectively). Overall, the results show that our proposed geometric constraint helps improve the photo-realism of generated images, as our enhanced images are consistently preferred over images from both the baseline model and ablation model.

6.4 Ablation Study

To evaluate the value of our proposed constraint, we perform extensive comparison between our enhanced model and the ablation model which was fine-tuned on the same dataset but without our

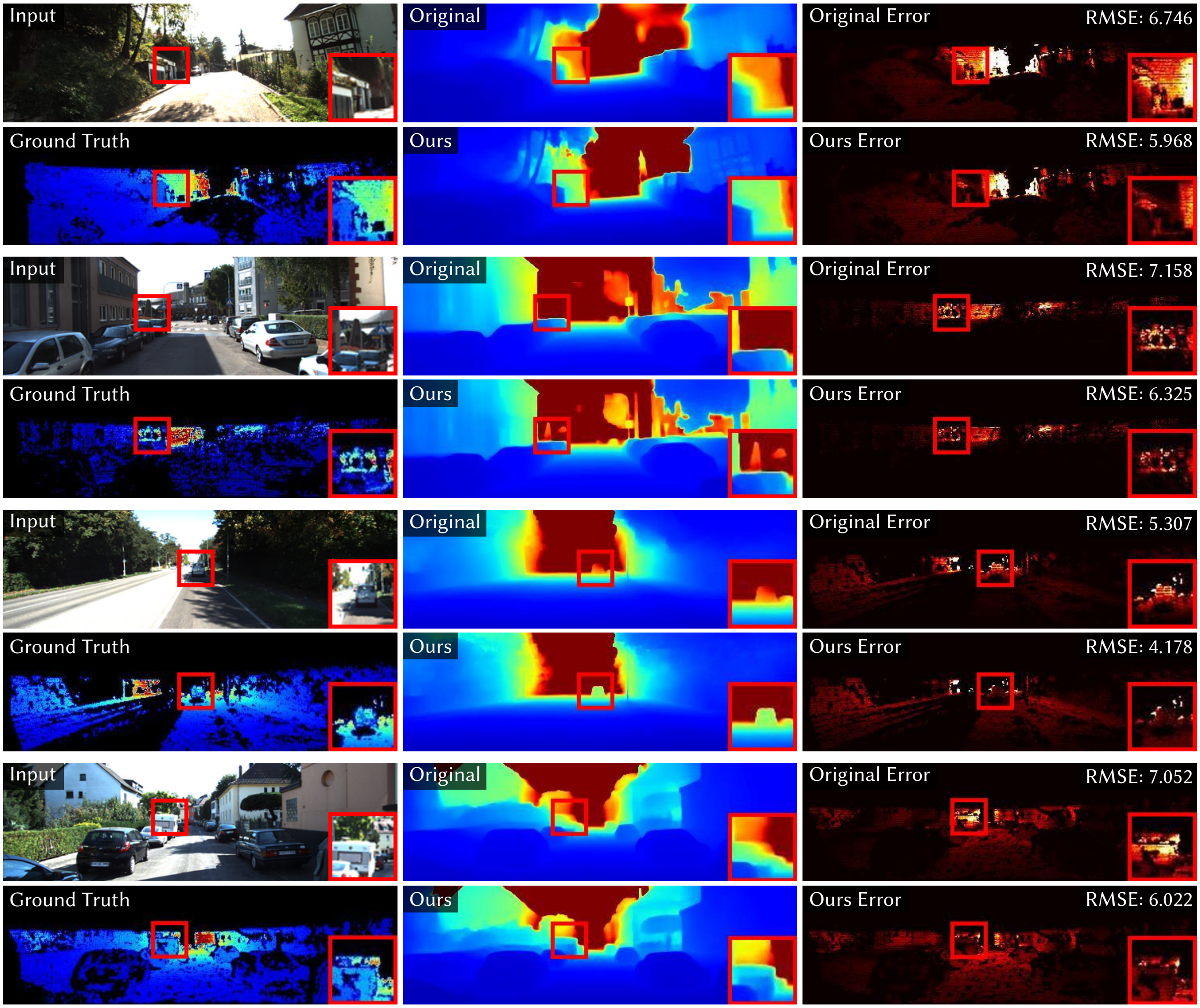


Fig. 9. **Qualitative comparisons of DPT-Hybrid fine-tuned on the data from our fine-tuned models and the original DPT-Hybrid model.** The depth maps produced by models trained on images from our enhanced model capture more high-frequency detail than the models trained on images from the baseline model. The RMSE error of the outputs of our model is also consistently lower.

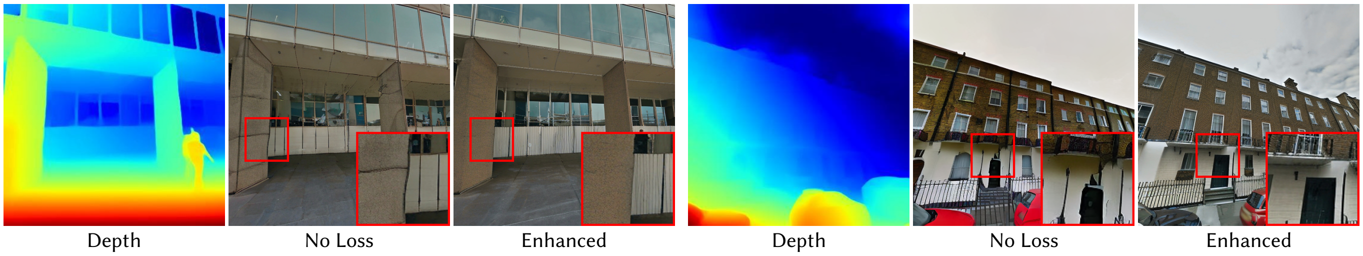


Fig. 10. **The proposed perspective constraint is responsible for the increase in perspective accuracy of generated images more than the dataset the diffusion models were fine-tuned on.** The depth maps these outputs are conditioned on are put in the left-hand column. Note that the images without our loss suffer from more distortions and curved lines and are less photo-realistic.

Table 1. **Monocular Depth Estimation performance of DPT-Hybrid fine-tuned on our data compared to the base DPT-Hybrid model.** The original DPT-Hybrid model was trained on a dataset referred to as MIX 6, which is a collection of 10 datasets as described in [Ranftl et al. 2021]. Fine-tuned models were trained on synthetic datasets generated by either the base stable diffusion model or our fine-tuned model. The best performing model is in bold and the second best is underlined.

Model	Description	Test Set	RMSE ↓	RMSE log ↓	AbsRel ↓	SqRel ↓	SiLog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DPT-Hybrid	Original	KITTI	5.0287	0.1874	0.1328	0.9705	18.6320	0.8385	0.9552	0.9855
	Fine-tuned on vKITTI Base		4.7680	0.1800	0.1286	0.8104	17.8890	0.8401	0.9587	0.9881
	Fine-tuned on vKITTI Enhanced		**4.6749**	**0.1760**	**0.1250**	**0.7827**	**17.4836**	**0.8496**	**0.9608**	**0.9890**
DPT-Hybrid	Original	DIODE Outdoor	9.5311	0.5667	0.4593	7.0644	52.6255	0.4709	0.6588	0.7759
	Fine-tuned on vKITTI Base		9.4863	0.5669	0.4560	**6.7930**	52.6316	0.4705	0.6586	0.7758
	Fine-tuned on vKITTI Enhanced		**9.4854**	**0.5663**	**0.4559**	6.8371	**52.5902**	**0.4713**	**0.6595**	**0.7763**

Table 2. **Monocular Depth Estimation performance of PixelFormer fine-tuned on our data compared to the base PixelFormer model (trained on KITTI) on the DIODE outdoor dataset.** Fine-tuned models were trained on synthetic datasets generated by either the base stable diffusion model or our fine-tuned model. The best performing model is in bold and the second best is underlined.

Model	Description	Test Set	RMSE ↓	RMSE log ↓	AbsRel ↓	SqRel ↓	SiLog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
PixelFormer	Original	DIODE Outdoor	8.8726	0.7041	1.4532	21.8911	66.0165	0.3254	0.5586	0.7075
	Fine-tuned on vKITTI Base		8.5381	0.6891	1.4140	21.8363	64.5891	0.3294	0.5651	0.7209
	Fine-tuned on vKITTI Enhanced		**8.4728**	**0.6870**	**1.3738**	**19.3406**	**64.4721**	**0.3329**	**0.5677**	**0.7245**
PixelFormer	Original	DIODE Outdoor	8.8726	0.7041	1.4532	21.8911	66.0165	0.3254	0.5586	0.7075
	Fine-tuned on All Base		8.5296	0.7109	1.4768	22.0467	66.6546	0.3270	0.5531	0.7038
	Fine-tuned on All Enhanced		**8.5109**	**0.7027**	**1.4408**	**21.5139**	**65.8426**	**0.3360**	**0.5635**	**0.7116**

Table 3. **Ablation Study: Monocular Depth Estimation performance of DPT-Hybrid fine-tuned on data from a model trained with no loss compared to the model trained with our loss.** The best performing model is in bold.

Model	Description	Test Set	RMSE ↓	RMSE log ↓	AbsRel ↓	SqRel ↓	SiLog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
DPT-Hybrid	Fine-tuned on vKITTI No Loss	KITTI	5.5733	0.2159	0.1573	1.1084	21.3919	0.7803	0.9389	0.9807
	Fine-tuned on vKITTI Enhanced		**4.6749**	**0.1760**	**0.1250**	**0.7827**	**17.4836**	**0.8496**	**0.9608**	**0.9890**
DPT-Hybrid	Fine-tuned on vKITTI No Loss	DIODE Outdoor	9.5241	0.5728	0.4573	**6.7422**	53.1904	0.4670	0.6581	0.7737
	Fine-tuned on vKITTI Enhanced		**9.4854**	**0.5663**	**0.4559**	6.8371	**52.5902**	**0.4713**	**0.6595**	**0.7763**
PixelFormer	Fine-tuned on vKITTI No Loss	DIODE Outdoor	8.5054	0.7047	1.3889	20.3750	66.5519	0.3184	0.5543	0.7035
	Fine-tuned on vKITTI Enhanced		**8.4728**	**0.6870**	**1.3738**	**19.3406**	**64.4721**	**0.3329**	**0.5677**	**0.7245**

Table 4. **Inpainting Quantitative Results: Images generated by our enhanced model out-perform both the baseline Stable Diffusion V2 model and Ablations on the LPIPS metric.** Our enhanced model performs best on all three datasets, while the ablation model is outperformed by the baseline model when tested on only landscapes. Lower is better for all columns.

Dataset	Holicity	Nature	All
# of Images	250	320	570
Baseline	0.1367	0.1584	0.1488
Ablation	0.1147	0.1659	0.1434
Ours	0.1138	0.1573	0.1382

proposed constraint. We include qualitative results in Fig. 10. The edges and corners of our images are more consistent than similar features in the baseline model's images. We also include quantitative comparisons between depth estimation models trained on the vKITTI dataset from our enhanced diffusion model and depth estimation models trained on the vKITTI dataset from our no loss diffusion model. The results from this experiment, for both DPT-Hybrid and PixelFormer, are shown in Table 4. The models trained on our enhanced model images outperform the models trained on the no loss model images on all metrics except for one (SqRel for DPT-Hybrid trained on the vKITTI dataset and tested on DIODE Outdoor). In addition, our model demonstrates significant improvements, up to 16.11% on RMSE, compared to the no loss model. These results demonstrate that the superior performance of downstream

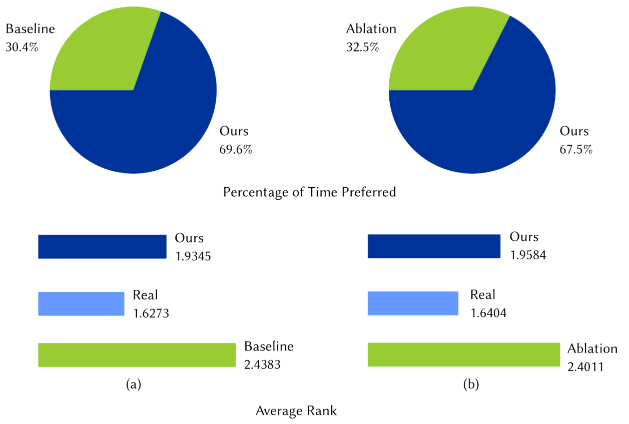


Fig. 11. **Images from our enhanced model consistently appear more photo-realistic than images from the baseline model (a) and our ablation model (b) according to the results of the subjective human tests.** *Top.* How often each set of images was ranked lower. Our enhanced images were ranked as more photo-realistic (lower) than baseline images in 69.6% of trials and were ranked as more photo-realistic than the ablation images in 67.5% of trials. *Bottom.* Average ranking for our images, real images, and comparison images. Although real images are consistently ranked the lowest, our images beat out both baseline and ablation images and are closer to real than the comparison.

models trained on our enhanced dataset is a result of our proposed constraint rather than a result of the new images introduced in fine-tuning. Beyond downstream tasks, the human subjective tests also show that our enhanced images are considered more photo-realistic than images from the ablation model 67.5% of the time (Fig. 11). In addition, quantitative and qualitative results (Fig. 7 and Table 4) on the inpainting task further highlight the improvement between our enhanced model and the ablation model. Combined, results from downstream tasks, human subjective tests, and the inpainting task demonstrate that the improvements achieved by our enhanced model are the result of our proposed geometric constraint rather than a result of fine-tuning on new images.

7 DISCUSSION

7.1 Limitations

One of the key limitations of our approach is that fine-tuning the diffusion model requires a dataset of images with vanishing points. However, these can be approximated using vanishing point detection tools [Lin et al. 2022; Liu et al. 2021]. Another limitation of our approach is the generation speed of latent diffusion models. On average, it takes ~3 seconds to generate a single image on 1 RTX3090, meaning generating a dataset of 150,000 images takes ~125 hours on 1 RTX3090. This significantly limits the potential size of synthetic datasets generated by latent diffusion models. Another limitation is that although our images are improved compared to the baseline model's images, they are still not quite at the level of real images as shown by our subjective test results. For example, Fig. 12 shows an image of Big Ben, and, although perspective lines are accurately depicted in the output, certain semantic details of the image are missing, the most obvious of which is that the clock is missing hands to show time. Other examples of this include nonsensical road signs or abstract store logos or flags.

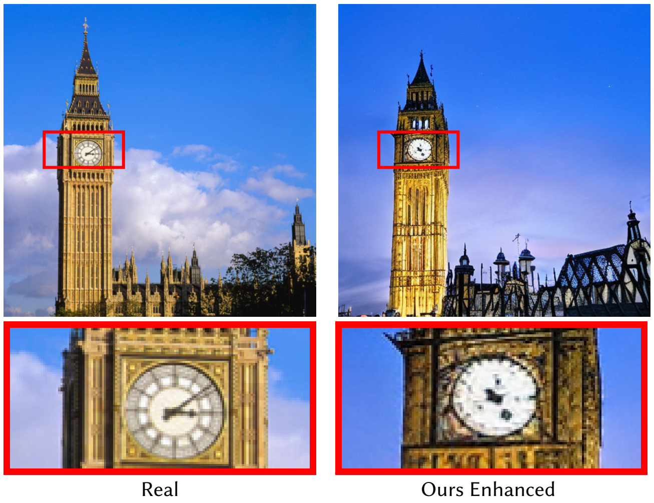


Fig. 12. **Outputs from stable diffusion are still unable to make certain semantic judgments.** Note that the clock shown on Big Ben is not functional and has no hour or minute hand.

7.2 Societal Impact

As always, there are downsides in improvements to generative models. As we increase the photo-realism of synthetic images, the potential for malicious use in the spread of disinformation also grows. In addition, perspective has been used as a tool to identify synthetic images from diffusion models [Corvi et al. 2022]. With the addition of our constraint, these tools could lose their efficacy, further increasing the potential for misuse of diffusion models.

7.3 Future Work

The current work is limited to 3D geometry perspective constraints, but there are still many other physical properties that affect the realism of generated images. One such example is lighting and shadow consistency [Farid 2022a,b] and semantic and physical consistency. Images generated by diffusion models often break physical laws, for example by having people walking on water. Future work can explore other constraints to help fulfill these physical laws and further increase photo-realism and the performance of downstream tasks.

7.4 Conclusions

In the 1400s, Leon Alberta Battisti established the foundations for perspective in art, which pushed the boundaries of hand-drawn realism. In this work, we propose a first attempt at a novel geometric constraint which encodes perspective into latent diffusion models. We demonstrate that introducing these physically-based 3D perspective constraints improves both photo-realism on subjective tests and downstream performance on monocular depth estimation. We hope that our work can be a small step in our community effort to improve the realism of image synthesis.

8 ACKNOWLEDGEMENTS

We thank members of the Visual Machines Group (VMG) at UCLA for feedback and support, as well as Guha Balakrishnan and Krish Kabra for technical discussions. A.K. was supported by a National Science Foundation (NSF) CAREER award IIS-2046737, Army Young Investigator Program Award, and Defense Advanced Research Projects Agency (DARPA) Young Faculty Award.

REFERENCES

Prolific Academic Ltd. 2023. Prolific · quickly find research participants you can trust. https://www.prolific.co/ Accessed on 2023-01-21.

Ashutosh Agarwal and Chetan Arora. 2023. Attention Attention Everywhere: Monocular Depth Prediction with Skip Attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5861–5870.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 214–223. https://proceedings.mlr.press/v70/arjovsky17a.html

Muhammad Awais. 2020. Animals dataset. https://www.kaggle.com/datasets/muhammadavici/animals-dataset

David Bau, Alex Andonian, Audrey Cui, YeonHwan Park, Ali Jahanian, Aude Oliva, and Antonio Torralba. 2021. Paint by word. *arXiv preprint arXiv:2103.10951* (2021).

Paul Beardsley and David Murray. 1992. Camera calibration using vanishing points. In *BMVC92*. Springer, 416–425.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=B1xsqj09Fm

Yohann Cabon, Naila Murray, and Martin Humenberger. 2020. Virtual KITTI 2. arXiv:2001.10773 [cs.CV]

Federico Camposeco and Marc Pollefeys. 2015. Using vanishing points to improve visual-inertial odometry. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 5219–5225. https://doi.org/10.1109/ICRA.2015.7139926

Jinming Cao, Oren Katzir, Peng Jiang, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. 2018. Dida: Disentangled synthesis for domain adaptation. *arXiv preprint arXiv:1805.08019* (2018).

Bruno Caprile and Vincent Torre. 1990. Using vanishing points for camera calibration. *International Journal of Computer Vision* 4 (1990), 127–139.

William Chen and Bernard C Jiang. 1991. 3-D camera calibration using vanishing point concept. *Pattern recognition* 24, 1 (1991), 57–67.

Rewon Child. 2021. Very Deep {VAE}s Generalize Autoregressive Models and Can Outperform Them on Images. In *International Conference on Learning Representations*. https://openreview.net/forum?id=RLRXCV6DbEJ

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2022. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680* (2022).

Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. 2018. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 447–463.

Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G Dimakis, and Peyman Milanfar. 2022. Soft diffusion: Score matching for general corruptions. *arXiv preprint arXiv:2209.05442* (2022).

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).

Hany Farid. 2022a. Lighting (In)consistency of Paint by Text. https://doi.org/10.48550/arxiv.2207.13744

Hany Farid. 2022b. Perspective (In)consistency of Paint by Text. https://doi.org/10.48550/arxiv.2206.14617

Xiaohan Fei, Alex Wong, and Stefano Soatto. 2019. Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters* 4, 2 (2019), 1661–1668.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4340–4349.

Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 2014. 3D Traffic Scene Understanding From Movable Platforms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 5 (2014), 1012–1025. https://doi.org/10.1109/TPAMI.2013.185

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)* (2013).

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

Andreas Georgis, Panagiotis Mermigkas, and Petros Maragos. 2022. VP-SLAM: A Monocular Real-time Visual SLAM with Points, Lines and Vanishing Points. *arXiv preprint arXiv:2210.12756* (2022).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Commun. ACM* 63, 11 (Oct 2020), 139–144. https://doi.org/10.1145/3422622

Erwan Guillou, Daniel Meneveaux, Eric Maisel, and Kadi Bouatouch. 2000. Using vanishing points for camera calibration and coarse 3D reconstruction from a single image. *The Visual Computer* 16, 7 (2000), 396–410.

Feng Han and Song-Chun Zhu. 2009. Bottom-Up/Top-Down Image Parsing with Attribute Grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1 (2009), 59–73. https://doi.org/10.1109/TPAMI.2008.65

BW He and YF Li. 2007. A novel method for camera calibration using vanishing points. In *2007 14th International Conference on Mechatronics and Machine Vision in Practice*. IEEE, 44–47.

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

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Achuta Kadambi. 2020. Blending physics with artificial intelligence. In *Computational Imaging V*, Vol. 11396. International Society for Optics and Photonics, 113960B.

Achuta Kadambi, Celso de Melo, Cho-Jui Hsieh, Mani Srivastava, and Stefano Soatto. 2023. Incorporating physics into data-driven computer vision. *Nature Machine Intelligence* (2023), 1–9.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114

Hui Kong, Jean-Yves Audibert, and Jean Ponce. 2009. Vanishing point detection for road detection. In *2009 ieee conference on computer vision and pattern recognition*. IEEE, 96–103.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 239–248.

Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326* (2019).

Young Hoon Lee, Changjoo Nam, Keon Yong Lee, Yuen Shang Li, Soo Yong Yeon, and Nakju Lett Doh. 2009. VPass: Algorithmic compass using vanishing points in indoor environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 936–941. https://doi.org/10.1109/IROS.2009.5354508

Bo Li, Kun Peng, Xianghua Ying, and Hongbin Zha. 2010. Simultaneous vanishing point detection and camera calibration from single images. In *International Symposium on Visual Computing*. Springer, 151–160.

Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. 2015. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 1119–1127.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086* (2022).

Hyunjun Lim, Jinwoo Jeon, and Hyun Myung. 2022. UV-SLAM: Unconstrained Line-Based SLAM Using Vanishing Points for Structural Mapping. *IEEE Robotics and Automation Letters* 7, 2 (2022), 1518–1525. https://doi.org/10.1109/LRA.2022.3140816

Yancong Lin, Ruben Wiersma, Silvia L Pintea, Klaus Hildebrandt, Elmar Eisemann, and Jan C van Gemert. 2022. Deep vanishing point detection: Geometric priors make dataset variations vanish. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6103–6113.

Shih-Ping Liou and Ramesh C Jain. 1987. Road following using vanishing points. *Computer vision, graphics, and image processing* 39, 1 (1987), 116–130.

Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. 2015. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence* 38, 10 (2015), 2024–2039.

Shichen Liu, Yichao Zhou, and Yajie Zhao. 2021. Vapid: A rapid vanishing point detector via learned optimizers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12859–12868.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition. 11461–11471.

Yi Ma, Stefano Soatto, Jana Kosecka, and S. Shankar Sastry. 2003. *An Invitation to 3-D Vision: From Images to Geometric Models.* SpringerVerlag.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. 2023. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 14297–14306.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which Training Methods for GANs do actually Converge?. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 3481–3490. https://proceedings.mlr.press/v80/mescheder18a.html

Danish Nazir, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. 2022. SemAttNet: Towards Attention-based Semantic Aware Guided Depth Completion. *IEEE Access* (2022), 1–1. https://doi.org/10.1109/ACCESS.2022.3214316

Justin Pinkney. 2022. stable-diffusion. https://github.com/justinpinkney/stable-diffusion.

Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. 2018. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 283–291.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning.* PMLR, 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 12179–12188.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* (2020).

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems* 32 (2019).

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

Robin Rombach, Andreas Blattmann, and Björn Ommer. 2022a. Text-guided synthesis of artistic images with retrieval-augmented diffusion models. *arXiv preprint arXiv:2207.13038* (2022).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention.* Springer, 234–241.

German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3234–3243.

Arnaud Rougetet. 2020. Landscape Pictures. https://www.kaggle.com/datasets/arnaud58/landscape-pictures

Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. 2018. Beyond sharing weights for deep domain adaptation. *IEEE transactions on pattern analysis and machine intelligence* 41, 4 (2018), 801–814.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).

Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 8503–8512.

Scott Satkin, Jason Lin, and Martial Hebert. 2012. Data-driven scene understanding from 3D models. (2012).

Ashutosh Saxena, Min Sun, and Andrew Y Ng. 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence* 31, 5 (2008), 824–840.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402* (2022).

Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. 2023. Depth Estimation from Camera Image and mmWave Radar Point Cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning.* PMLR, 2256–2265.

Arash Vahdat and Jan Kautz. 2020. NVAE: A Deep Hierarchical Variational Autoencoder. In *Neural Information Processing Systems (NeurIPS).*

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR* abs/1908.00463 (2019). http://arxiv.org/abs/1908.00463

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. 2023. End-to-End Diffusion Latent Optimization Improves Classifier Guidance. *arXiv preprint arXiv:2303.13703* (2023).

Bram Wallace, Akash Gokul, and Nikhil Naik. 2022. EDICT: Exact Diffusion Inversion via Coupled Transformations. *arXiv preprint arXiv:2211.12446* (2022).

Guanghui Wang, Hung-Tat Tsui, and Q. M. Jonathan Wu. 2009. What can we learn about the scene structure from three orthogonal vanishing points in images. *Pattern Recognit. Lett.* 30 (2009), 192–202.

Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. 2022. Synthetic Generation of Face Videos with Plethysmograph Physiology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 20587–20596.

Alex Wong, Safa Cicek, and Stefano Soatto. 2021. Learning topology from synthetic data for unsupervised depth completion. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1495–1502.

Alex Wong and Stefano Soatto. 2019. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5644–5653.

Alex Wong and Stefano Soatto. 2021. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 12747–12756.

Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 3917–3925.

Fengting Yang and Zihan Zhou. 2018. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV).* 85–100.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* (2022).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR.*

Yichao Zhou, Jingwei Huang, Xili Dai, Linjie Luo, Zhili Chen, and Yi Ma. 2020. HoliCity: A City-Scale Data Platform for Learning Holistic 3D Structures. (2020). arXiv:2008.03286 [cs.CV].

Javad Zolfaghari Bengar, Abel Gonzalez-Garcia, Gabriel Villalonga, Bogdan Raducanu, Hamed H Aghdam, Mikhail Mozerov, Antonio M Lopez, and Joost van de Weijer. 2019. Temporal Coherence for Active Learning in Videos. *arXiv preprint arXiv:1908.11757* (2019).