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

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. 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:

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.
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 perspective accurately 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 baseline 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
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
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 in image space and vanishing points in image space 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.

Natural Images. By the math of perspective projection for a pinhole camera, an point \( X = (X, Y, Z) \) is projected to a point \( x = (x, y) = (fX/Z, fY/Z) \) [Ma et al. 2003]. If we are concerned with a line \( L = O + rD \), 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 \( D \). 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 [Pilkney 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 \( x_0, \ldots, x_T \):

\[
q(x_t | x_{t-1}) \equiv \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)I),
\]

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

\[
p(x_{t-1} | x_t) \equiv \mathcal{N}(\mu_{\theta}(x_t, t), \Sigma(x_t, t)),
\]

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

\[
\mu_{\theta}(x_t, t) = \frac{1}{\sqrt{\beta_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon(x_t, t) \right).
\]

where \( \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \) and \( \epsilon(x_t, t) \) is a learned function parameterized by a UNet model [Ronneberger et al. 2015] with learned parameters \( \theta \). Based on this, the traditional diffusion model loss is as follows:

\[
L_{DM} = \mathbb{E}_{x_t \sim \mathcal{N}(0, I), t} \left[ \| I - \epsilon_o(x_t, t) \|^2_2 \right].
\]

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: \( E : X \in \mathbb{R}^{H \times W \times 3} \rightarrow Z \in \mathbb{R}^{H' \times W' \times 3} \), while the decoder is defined as: \( D : Z \in \mathbb{R}^{H' \times W' \times 3} \rightarrow X \in \mathbb{R}^{H \times W \times 3} \), where \( h = H/f, w = W/f \) and \( f \) is a downsampling factor. With this formulation, the loss function now becomes:

\[
L_{DM}^\prime = \mathbb{E}_{(x, z) \sim \mathcal{D}(0,1), t} \left[ \| I - \epsilon_o(z_t, t) \|^2_2 \right],
\]

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:

\[
L_{DM}^\prime = \mathbb{E}_{(x, z) \sim \mathcal{D}(0,1), t} \left( \| I - \epsilon_o(z_t, t) \|^2_2 \right) + \lambda \mathbb{E}_{x \sim \mathcal{D}(0,1), t} \left[ L_{\text{persp}}(\hat{x}, x, v_x) \right],
\]

where \( \lambda \) is a weight factor for our perspective loss, \( v_x \) is a set of vanishing points in image space and \( \hat{x} \) is our reconstructed image, which can be written as:

\[
\hat{x} = D \left( \frac{1}{\sqrt{\beta_t}} \left( x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon(x_t, t) \right) \right).
\]

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

- \( G_x \) represents the gradients of an image \( x \) computed with a 3x3 Sobel filter.

ACM Trans. Graph., Vol. 1, No. 1, Article . Publication date: August 2023.
we detail how we fine-tune monocular depth estimation models on images generated from our fine-tuned models. In Section 5.3, we conduct comprehensive experiments. In Section 5.1, we detail how we fine-tune monocular depth estimation models. In particular, we fine-tune the baseline model to the loss function detailed in Eq. 6. 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 has saturated. All models were trained on 4 RTX3090 GPUs. Results are shown in Section 6.1.

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 captions [Schuhmann et al. 2022]. In this paper, we refer to this model as the baseline model.

Datasets. 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. 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 512x512 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 has 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. 2015].

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.

```plaintext
ALGORITHM 1: Algorithm to compute perspective loss

Function perspective_loss(x, x̂, v_k)

Input : Image x
Input : Ground Truth image x̂
Input : Vanishing Points v_k

G_k ← img_derivative(x)
G_k ← img_derivative(x̂)
loss ← 0.0

foreach v ∈ v_k do
    φ_min, φ_max ← calc_image_angle(v)
    for i ← 0; i < N; i = i + 1 do
        angle ← i/N * (φ_max - φ_min) + φ_min
        d ← calc_perp_vec(angle)
        p ← get_line_pixels(v, angle)
        D(i) ← Σ_p |G_k · d|
        D0(i) ← Σ_p |G_k · d|
        loss ← loss + norm(D - D0)
    end
    loss ← loss/|v|
end
return loss
```

- $φ_{\text{min}}$ and $φ_{\text{max}}$ 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 $φ_{\text{min}}$ and $φ_{\text{max}}$.
- $v$ represents a particular vanishing point in the set $v_k$.
- $l_i(v, k)$ 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:

$$D_i(v, x) = \int_{k_0}^{k_1} d_i(v) · G_k(l_i(v, k)) dk,$$  (8)

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

$$L_{\text{perp}}(x, x̂, v_k) = \frac{1}{|v_k|} \sum_{v \in v_k} \|D(v, x) - D(v, x̂)\|_2.$$  (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.
we use the original images, resulting in a total of 70,000 images. All Enhanced refers to the full set of 155k images generated by (Real, Baseline, Ours or Real, Ablation, Ours) in order of photo-

vKITTI, and refer to this dataset as vKITTI. We additionally append 1.25, 1

Human Subjective Test Methodology

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 ($\delta_i$) at thresholds $\tau_i = 1.25, 1.25^2, 1.25^3$ 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.

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.
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].

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.
6.1 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 [Rouget et al. 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 five out of eight metrics, but outperforms our model on no metrics. In particular, our model shows a 7.03% improvement in RMSE 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 3x3 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.

ACM Trans. Graph., Vol. 1, No. 1, Article . Publication date: August 2023.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Test Set</th>
<th>RMSE</th>
<th>RMSE log</th>
<th>AbsRel</th>
<th>SqRel</th>
<th>SiLog</th>
<th>𝛿1↑</th>
<th>𝛿2↑</th>
<th>𝛿3↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPT-Hybrid</td>
<td>Original</td>
<td>KITTI</td>
<td>5.573</td>
<td>0.1874</td>
<td>0.1328</td>
<td>0.9705</td>
<td>18.6320</td>
<td>0.8385</td>
<td>0.9552</td>
<td>0.9855</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Base</td>
<td>KITTI</td>
<td>4.7689</td>
<td>0.1800</td>
<td>0.1286</td>
<td>0.8104</td>
<td>17.8890</td>
<td>0.8401</td>
<td>0.9587</td>
<td>0.9881</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Enhanced</td>
<td>KITTI</td>
<td>4.6749</td>
<td>0.1760</td>
<td>0.1250</td>
<td>0.7827</td>
<td>17.4836</td>
<td>0.8496</td>
<td>0.9608</td>
<td>0.9890</td>
</tr>
<tr>
<td>DPT-Hybrid</td>
<td>Original</td>
<td>DIODE Outdoor</td>
<td>9.5311</td>
<td>0.5667</td>
<td>0.4593</td>
<td>7.0644</td>
<td>52.6255</td>
<td>0.4709</td>
<td>0.6588</td>
<td>0.7759</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Base</td>
<td>DIODE Outdoor</td>
<td>9.4863</td>
<td>0.5669</td>
<td>0.4560</td>
<td>6.7930</td>
<td>52.6316</td>
<td>0.4705</td>
<td>0.6586</td>
<td>0.7758</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Enhanced</td>
<td>DIODE Outdoor</td>
<td>9.1454</td>
<td>0.5663</td>
<td>0.4559</td>
<td>6.8371</td>
<td>52.5902</td>
<td>0.4713</td>
<td>0.6595</td>
<td>0.7763</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Test Set</th>
<th>RMSE</th>
<th>RMSE log</th>
<th>AbsRel</th>
<th>SqRel</th>
<th>SiLog</th>
<th>𝛿1↑</th>
<th>𝛿2↑</th>
<th>𝛿3↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>PixelFormer</td>
<td>Original</td>
<td>DIODE Outdoor</td>
<td>8.8726</td>
<td>0.7041</td>
<td>1.4532</td>
<td>21.8911</td>
<td>66.0165</td>
<td>0.3254</td>
<td>0.5586</td>
<td>0.7075</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Base</td>
<td>DIODE Outdoor</td>
<td>8.5381</td>
<td>0.6981</td>
<td>1.4140</td>
<td>21.8363</td>
<td>64.5891</td>
<td>0.3294</td>
<td>0.5651</td>
<td>0.7209</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Enhanced</td>
<td>DIODE Outdoor</td>
<td>8.4728</td>
<td>0.6870</td>
<td>1.3738</td>
<td>19.3406</td>
<td>64.4721</td>
<td>0.3329</td>
<td>0.5677</td>
<td>0.7245</td>
</tr>
<tr>
<td>PixelFormer</td>
<td>Original</td>
<td>DIODE Outdoor</td>
<td>8.8726</td>
<td>0.7041</td>
<td>1.4532</td>
<td>21.8911</td>
<td>66.0165</td>
<td>0.3254</td>
<td>0.5586</td>
<td>0.7075</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on All Base</td>
<td>DIODE Outdoor</td>
<td>8.5296</td>
<td>0.7109</td>
<td>1.4768</td>
<td>22.0467</td>
<td>66.6546</td>
<td>0.3270</td>
<td>0.5551</td>
<td>0.7058</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on All Enhanced</td>
<td>DIODE Outdoor</td>
<td>8.5109</td>
<td>0.7027</td>
<td>1.4408</td>
<td>21.5139</td>
<td>65.8426</td>
<td>0.3360</td>
<td>0.5635</td>
<td>0.7116</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Test Set</th>
<th>RMSE</th>
<th>RMSE log</th>
<th>AbsRel</th>
<th>SqRel</th>
<th>SiLog</th>
<th>𝛿1↑</th>
<th>𝛿2↑</th>
<th>𝛿3↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPT-Hybrid</td>
<td>Fine-tuned on vKITTI No Loss</td>
<td>KITTI</td>
<td>5.5733</td>
<td>0.2159</td>
<td>0.1573</td>
<td>1.1084</td>
<td>21.919</td>
<td>0.7803</td>
<td>0.9389</td>
<td>0.9807</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Enhanced</td>
<td>KITTI</td>
<td>4.6749</td>
<td>0.1760</td>
<td>0.1250</td>
<td>0.7827</td>
<td>17.4836</td>
<td>0.8496</td>
<td>0.9608</td>
<td>0.9890</td>
</tr>
<tr>
<td>DPT-Hybrid</td>
<td>Fine-tuned on vKITTI No Loss</td>
<td>DIODE Outdoor</td>
<td>9.5241</td>
<td>0.5728</td>
<td>0.4573</td>
<td>6.7422</td>
<td>53.1904</td>
<td>0.4670</td>
<td>0.6581</td>
<td>0.7737</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Enhanced</td>
<td>DIODE Outdoor</td>
<td>9.4854</td>
<td>0.5663</td>
<td>0.4559</td>
<td>6.8371</td>
<td>52.5902</td>
<td>0.4713</td>
<td>0.6595</td>
<td>0.7763</td>
</tr>
<tr>
<td>PixelFormer</td>
<td>Fine-tuned on vKITTI No Loss</td>
<td>DIODE Outdoor</td>
<td>8.5054</td>
<td>0.7027</td>
<td>1.3889</td>
<td>20.3750</td>
<td>66.5519</td>
<td>0.3184</td>
<td>0.5543</td>
<td>0.7035</td>
</tr>
<tr>
<td></td>
<td>Fine-tuned on vKITTI Enhanced</td>
<td>DIODE Outdoor</td>
<td>8.4728</td>
<td>0.6870</td>
<td>1.3738</td>
<td>19.3406</td>
<td>64.4721</td>
<td>0.3329</td>
<td>0.5677</td>
<td>0.7245</td>
</tr>
</tbody>
</table>

Table 4. Inpainting Quantitative Results: Images generated by our enhanced model outperform 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Holecity</th>
<th>Nature</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Images</td>
<td>250</td>
<td>320</td>
<td>570</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.1367</td>
<td>0.1584</td>
<td>0.1488</td>
</tr>
<tr>
<td>Ablation</td>
<td>0.1147</td>
<td>0.1659</td>
<td>0.1434</td>
</tr>
<tr>
<td>Ours</td>
<td>0.1138</td>
<td>0.1573</td>
<td>0.1382</td>
</tr>
</tbody>
</table>

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.

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.
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


Enhancing Diffusion Models with 3D Perspective Geometry Constraints


