

Towards Visual Foundation Models of Physical Scenes

Chethan Parameshwara* Alessandro Achille* Matthew Trager Xiaolong Li
 Jiawei Mo Jianbo Ye Ashwin Swaminathan C.J. Taylor
 Dheera Venkatraman Xiaohan Fei* Stefano Soatto*
 {xiaohfei,soattos,aachille,cmparam}@amazon.com

AWS AI Labs
 May 17, 2023

Abstract

We describe a first step towards learning general-purpose visual representations of physical scenes using only image prediction as a training criterion. To do so, we first define “physical scene” and show that, even though different agents may maintain different representations of the same scene, there is a notion of physical scene that can be uniquely defined and inferred. Then, we show that NeRFs cannot represent the physical scene, as they lack extrapolation mechanisms. Those, however, could be provided by Diffusion Models, at least in theory. To test this hypothesis empirically, NeRFs can be combined with Diffusion Models, a process we refer to as NeRF Diffusion, used as unsupervised representations of the physical scene. Our analysis is limited to visual data, without external grounding mechanisms that can be provided by independent sensory modalities.

1 Introduction

Vision serves to infer properties of *the scene* from images. But in reality each agent, whether natural or artificial, can only perceive the scene through a finite set of observations, which are compatible with infinitely many scenes. It is not clear, then, what “*the scene*” even means, and how scenes inferred by different agents for different purposes may relate. For example, for the purpose of navigating the surrounding environment, the scene may be fruitfully described as a geometric configuration of surfaces, whereas for the purpose of recognizing edible objects, photometric statistics aggregated locally may be more useful. In general, different tasks may require inferring geometric (shape), photometric (reflectance, illumination), dynamic (motion), functional (affordances) or semantic (identities and relations) properties of the scene.

It seems implausible, therefore, that models trained to simply predict images, such as Neural Radiance Fields (NeRFs) and Diffusion Models (DMs), oblivious of the complexities of light and matter, may “capture the scene” in the sense of inferring a representation that can subtend all visual tasks. However, data prediction is a universal task and, in other domains such as language modeling, simply predicting unseen tokens appears sufficient to learn a representation that can support a multitude of downstream tasks and seemingly engender high-level behavior such as reasoning (“chain-of-thought”) and transduction (“in-context learning”). At first glance, the language realm and physical environment appear rather different: Language originates in the human brain which is not accessible, whereas the physical environment can be probed with multiple independent modalities. Yet, we argue, the “true scene” is a chimera, and there may be more similarities between “language models” and “world models” than meet the eye.

In this paper, we explore whether and how NeRFs and/or Diffusion Models *can* learn a representation of *the scene*. The outline of the paper and its contributions are summarized next.

*Equal contribution

High-level Summary

First, we (i) define the scene in Sect. 2.1. We argue that, from the point of view of an agent that makes observations, the scene can only be meaningfully defined as an *abstract concept*, with no underlying objective entity. Nonetheless, (ii) if a “true” or “objective” scene exists which generates the measurements, we show that it fits our definition (Theorem 2.5). Unfortunately, (iii) this is not useful since, given two valid scenes compatible with the measurements, an agent cannot decide which of the two, if any, is the “real” one (Theorem 2.6). Nonetheless, we can (vi) define a notion of “physical scene” that is unique and therefore unambiguous (Theorem 2.10).

The results in Section 2 suggests that an abstract model, such as that implemented by a Deep Neural Network, can in principle represent the physical scene. NeRFs are an obvious first candidate, since ostensibly they are designed and trained to infer the radiance function, whose structure depends on the geometry and photometry of the scene. Unfortunately, (vii) in Section 3 we show that a NeRF, at least in its basic implementations, *cannot* represent a scene (Proposition 3.1). However, Diffusion Models provide what is missing: In Proposition 3.2 we show that (viii) the composition of a NeRF with a Diffusion Model, which we refer to as NeRF Diffusion, can be a viable representation of the physical scene.¹

The argument summarized in (i)-(viii) represents the basis on which we design our first implementation of a model that, potentially, can represent physical scenes. The conditions under which a model can represent a scene require that the model can not only interpolate, but also *extrapolate* details that are not manifest in a dataset, doing so in a manner that is compatible with the aggregate statistics of all seen scenes. To empirically validate the claim that a NeRF augmented with a Diffusion Model is a valid representation, we therefore need to show that it can “hallucinate” unseen details in a way that is perceptually¹ if not objectively, accurate.

To this end, in Section 4 we describe (ix) a novel diffusion-based method to learn the structural priors and sampling operators so that, composed with a NeRF, a Diffusion Model can extrapolate realistic images. Instead of training diffusion models from scratch, we fine-tune pre-trained Stable Diffusion with NeRF rendering to shape the priors. Through learned priors, the diffusion model enhances the rendering on novel view synthesis to the point where it could be used to extrapolate details indefinitely.

To the best of our knowledge, we are first to attempt a derivation of the characteristics that a trained model should have in order to represent “the physical scene” without any objectivity requirement, using data alone. In this sense, little prior work in the literature has tackled (i)-(viii) outside the language domain [27]. There is, on the other hand, a massive and rapidly growing literature on NeRFs and Diffusion Models, as well as on their combination (ix), which we point to throughout the paper and describe in a Sec. 6 in order to not disrupt the flow. In Sec. 7 we discuss the limitations of our approach and potential avenues for future work.

2 Background: Representations of Visual Scenes

We think of a camera as a function that yields measurements that depend on its own configuration as well as on a latent entity, or “scene.” As anticipated, defining the scene as an objective entity is problematic, but we posit that, whatever it may be, it is shared among all agents immersed in it. Since each agent observes different views of it, or even if they all observe the same views, each may process them differently or impose different priors, the representations of the shared scene are different in each agent; yet, all these versions are *related* in the sense of being representations of the same underlying scene. Since no version is special in any sense, what defines the scene, then, is none of them, but rather their relation, as we articulate in the next subsections.

¹This is the model’s perception, not human perception: The criterion is for synthesized and real images to be phenomenologically indistinguishable by the model itself.

2.1 What is “a scene”? Scenes as abstract concepts

Let $y = (y_1, \dots, y_k) \in \mathbb{R}^k$ be a random variable representing the measurements coming from k different sensors. Associated to these measurements, there is a configuration $c \in \mathcal{C}$ of the sensors (e.g., the pose of the camera). To simplify the notation, we write $\mathbf{y} = (y, c)$. An agent observes samples from the joint distribution $p(y, c)$.² After several measurements are obtained, $(\mathbf{y}_1, \dots, \mathbf{y}_n)$, the agent may observe that they are correlated. For example, moving around a table we may observe that the *perceived* brown quadrilateral maintains an overall similar color and its shape changes predictably with the viewpoint c [33]. These correlations could be “explained” by an underlying factor, which we call *a scene*:

Definition 2.1 (A Scene). A scene is any random variable $S \in \mathcal{S}$ that renders the measurements independent, that is

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n) = \int p(\mathbf{y}_1|S) \dots p(\mathbf{y}_n|S) dP(S).$$

In particular, no pair of measurements share any information $I(\mathbf{y}_1; \mathbf{y}_2|S) = 0$, once a scene is known. In this sense, a scene “explains” the measurements.³ We call *representation of a scene* the pair $\mathbf{r} = (p(\mathbf{y}|S), P(S))$ of probability distributions that define the scene.

Remark 2.2 (Representation and hallucination). We wish to emphasize the difference between a representation of a scene and a representation of the given images. For a finite set of observations, $\mathbf{y}_{\leq t} = (\mathbf{y}_1, \dots, \mathbf{y}_t)$, a representation is simply any sufficient statistic, including the data themselves. On the other hand, a representation of the scene comprises two elements: $p(\mathbf{y}|S)$, which can be used to generate observations from the scene, and $P(S)$ which can be seen as a *prior* over the possible scenes. A non-trivial consequence of this choice is that a representation of the scene can be used to *hallucinate*, or predict realistic measurements, something that a representation of the measurements does not. To see this, consider the problem of hallucinating a new measurement $\mathbf{y} \sim p(\mathbf{y}|\mathbf{y}_{\leq t})$ conditioned on past observations. This can be written as:

$$p(\mathbf{y}|\mathbf{y}_{\leq t}) = \frac{p(\mathbf{y}, \mathbf{y}_{\leq t})}{p(\mathbf{y}_{\leq t})} = \int p(\mathbf{y}|S) \frac{p(\mathbf{y}_1|S) \dots p(\mathbf{y}_t|S)}{\int p(\mathbf{y}_1|S) \dots p(\mathbf{y}_t|S) dP(S)} dP(S) = \int p(\mathbf{y}|S) dP(S|\mathbf{y}_{\leq t}).$$

From the second equality, we see that knowing the scene representation $\mathbf{r} = (p(\mathbf{y}|S), P(S))$ gives us all the information we need to hallucinate new realistic observations.

Based on this definition, “the scene” is not unique, which is why we refer to it as “a scene” instead. Later we will show that, if “*the*” scene exists, in the sense commonly referred to as the “real” or “true” scene, then it is a scene in the sense of the definition. But, while the agent cannot know whether such a “true” scene actually exists, a scene always does, under mild assumptions:

Theorem 2.3 (Existence of a scene). Suppose the measurements are exchangeable, meaning that

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n) = p(\mathbf{y}_{\pi(1)}, \dots, \mathbf{y}_{\pi(n)}),$$

where $\pi : [n] \rightarrow [n]$ is any permutation of $[n] = \{1, \dots, n\}$ and that each finite sequence can be extended to an infinite exchangeable sequence of measurements. Then, there exists a random variable S such that

$$p(\mathbf{y}_1, \dots, \mathbf{y}_n) = \int p(\mathbf{y}_1|S) \dots p(\mathbf{y}_n|S) dP(S),$$

hence S is a scene.

²It could be argued that agents do not observe the configuration c of their own sensors but either control it, or have to infer it from the data. However, taking this point of view requires introducing a notion of temporal continuity which complicates the theory. Simply assuming that c is known or already inferred from data allows us to develop most of the theory without having to deal with time, and without loss of generality since whatever part of c cannot be inferred from y does not affect the representation anyway.

³The reader may be wondering why we are integrating with respect to the scene, when a NeRF is inferred from measurements of a single scene. Unfortunately, the language of statistics does not allow us to deal with individual entities, but one can imagine the agent being dropped into random scenes, each time executing a number of measurement, and repeating the process a number of times without knowledge of whether the underlying scene is the same, indefinitely.

This follows directly from De Finetti's theorem [9]. Now that we have (at least) a scene, we define a measurement function, or *presentation* [15] *not* of the real scene, but of any scene.

Definition 2.4 (Measurement function and “presentation”). Let $S \in \mathcal{S}$ be a scene. The induced measurement function h is a stochastic function $h : \mathcal{S} \times \mathcal{C} \rightarrow \mathbb{R}^k$ defined by $h(S, c) = y$ where $y \sim p(y|S, c)$. We call the function $p(y|\cdot, c)$ a *presentation*, which is instantiated for any given scene as $p(y|S, c)$.

Note that specifying the presentation function, as a computational procedure, is equivalent to specifying a scene, although we do not refer to as a *representation* of the scene, for reasons clarified in Remark 2.13. The fact that specifying a scene and its presentation are equivalent should make it clear that a scene, as defined, is not some “objective,” “true,” “material,” or “physical reality,” but an entirely abstract entity that may be embodied in a number of ways, for instance as a digital computer or as a neural network. By sampling from $p(\cdot|S, c)$, a scene can produce infinitely many “controlled hallucinations” [15], as discussed in detail in Remark 2.13. While not objective, a scene is connected to reality in two ways: In one direction, if the real scene exists, then it is a scene, as we show next. In the other direction, if all scenes can be related to each other in some canonical way, then we can define such a canonical scene as a proxy of the “real” scene without any ontological complications. We do so in the next section.

Theorem 2.5 (The “real scene”, if it exists, is a scene). Assume that there is an objective entity S that generates the measurement of the agent thorough a measurement function h . That is, assume that $y = h(S, c)$. Then S is a scene.

This follows directly from the definition but, despite being seemingly innocuous, the statement is problematic since any passive observer, whether natural or artificial, has only access to the given measurements, which are never enough to capture the “real” scene unless we make strong assumptions about it (for instance finiteness), which would inevitably be unverifiable [1]. This is captured by the next statement.

Claim 2.6 (A passive agent cannot know whether the scene is real). Let S and S' be two scenes. Suppose that the sequence of measurements $\mathbf{y}_1, \dots, \mathbf{y}_n$ is generated by first flipping a fair coin $z \sim \text{Bernoulli}(\frac{1}{2})$ to decide whether to use S or S' , and then sampling measurements from that scene using its respective measurement function. Then the agent cannot infer z beyond chance level.

This is a simple consequence of the fact that the agent only observes samples from $p(\mathbf{y}_1, \dots, \mathbf{y}_n)$ and both scenes have the same marginal distribution. The upshot is that, if there are two scenes that generate identical measurements, an agent cannot tell the difference, so the notion of *the “true scene”* is a chimera. Moreover, two scenes S and S' may be incompatible, meaning that there may be no computable functions f or g such that $S = f(S')$ or $S' = g(S)$. Hence, even if there was a real scene underlying the data, each individual agent may infer its own version, which bears no relation to that of another agent, resulting in a perceptual Babel. The question, then, is whether there is some notion of scene that all agents who share the same observations can agree to, so they can at least share information about it. We call this the *physical scene*. We preface that, by necessity, the physical scene is still an abstract concept, dependent on the measurements available to each agent.

Since individual scenes are not directly comparable, we therefore seek some sort of minimality criterion that each agent can enforce in order to arrive at some sort of “canonical” scene which, if unique in some sense, we can refer to as the *physical scene*.

2.2 How are scenes related? “The physical scene”

In Computer Vision we tend to think of “the scene” as the underlying “cause” of a given collection of images. Unfortunately, as we saw in the previous section, this *naive realism* is fallacious⁴ [10]. Instead, we flip the

⁴Quoting (pp. 14-15): “We all start from ‘naive realism’, i.e., the doctrine that things are what they seem. [...] The observer, when he seems to himself to be observing a stone, is really, if physics is to be believed, observing the effects of the stone upon himself. Thus science seems to be at war with itself: when it most means to be objective, it finds itself plunged into subjectivity against its will. Naive realism leads to physics, and physics, if true, shows that naive realism is false. Therefore naive realism,

script and think of the given images, which are objective, as underlying an infinite set of possible scenes, and the key question is how all these scenes are *related*.

Definition 2.7 (Sufficient statistics of a scene). A sufficient statistic of S for the measurement is any function $T(S)$ such that

$$p(\mathbf{y}|S) = p(\mathbf{y}|T(S)),$$

or equivalently such that we have the Markov chain $S \rightarrow T(S) \rightarrow y$. Note that this is a sufficient statistic of *the scene* to generate the measurement, not a sufficient statistic of the measurement to infer the scene.⁵ Equivalently, $T(S)$ is a subset of the information in the scene that is sufficient to explain the measurement $I(\mathbf{y}; S) = I(\mathbf{y}; T(S))$.

Note that thus far we have said nothing about how a sufficient scene S arises. All we have said is that, if an entity exists that satisfies the definition of a scene, it can generate all the measurements that the real scene, if it existed, would generate. Later we will address identifiability and observability. For now, of all sufficient statistics, we are interested in the simplest possible ones, which are generally not unique.

Definition 2.8 (Minimal sufficient statistics). A minimal sufficient statistic $T(S)$ of S is a sufficient statistic such that, given any other sufficient statistic $T'(S)$ we have $T(S) = f(T'(S))$ for some function f .

We now have a candidate for a notion of scene that agents who observe the same measurements could share. Next, we will prove that such a scene is unique in a strong sense, making it canonical, which allows us to refer to it as “*the (physical) scene*”.

Theorem 2.9 (Existence of a minimal sufficient scene). Given a sufficient scene S , we can always construct a minimal sufficient scene $S_m = T(S)$ under the weak condition that the support of $p(S|\mathbf{y})$ is the same for all \mathbf{y} .⁶ In particular, the function $L : \mathcal{S} \times \mathcal{M} \rightarrow \mathbb{R}$ defined by

$$L(S, y) = \frac{p(S|\mathbf{y})}{p(S|\mathbf{y}_0)} \propto \frac{p(\mathbf{y}|S)}{p(\mathbf{y}_0|S)}$$

for any $\mathbf{y}_0 \in \mathcal{M}$ is a minimal sufficient scene.

The following result justifies naming a minimal sufficient scene *the physical scene*.

Theorem 2.10 (Strong uniqueness and the physical scene). Let S and S' be two sufficient scenes and let S_y and S'_y be two minimal sufficient scenes corresponding to S and S' respectively. Then there are functions f and g such that $S_y = f(S'_y)$ and $S'_y = g(S_y)$.

We therefore call the unique minimal sufficient scene for a set of given measurements the *physical scene* subtending those measurements. Note that the physical scene has to be compatible with physical laws that involve measured quantities, for such laws could be interpreted as production rules for the measurements. Both physical laws and physical scenes are abstract concepts, in the sense that they can be represented in the memory of the agent, even though we have not yet described how such a scene could be inferred from data.

The existence theorem is a straightforward application of the standard existence theorem for minimal sufficient statistics [14, 5]. The uniqueness theorem, however, is stronger, as normally we would conclude that all minimal sufficient statistics of the same scene are in a bijection, but could not compare between different scenes. To get the stronger version, note that if S and S' are scenes, then the tuple (S, S') is also a scene and both S_y and S'_y are minimal sufficient statistics of it. It then follows that they are in a bijection. Since a bijection is an equivalence relation, we can also represent the physical scene as an equivalence class, drawing a parallel to concepts in natural language, as we describe next.

if true, is false; therefore it is false.

⁵This flipping of the focus from the data to the scene corresponds to a change in perspective from *naive realism* to a *analytical philosophy* or, using Koenderink’s nomenclature [15], from the “Marrian” to the “Goethean” accounts of Vision.

⁶This is the case, for instance, if the measurements are affected by Gaussian noise, whose support covers the domain.

Remark 2.11 (Scenes as equivalence classes). As we anticipated, any given collection of images is compatible with infinitely many scenes, none of which is “special” (canonical), so the physical scene cannot be any one of them, but rather their *relation*, which is a bijection, hence an *equivalence relation*. Specifically, let S be a scene and $p(y|S, c)$ its *presentation*. Then, presentations define an equivalence relation among scenes, whereby $S' \sim S \Leftrightarrow p(y|S, c) = p(y|S', c) \forall c$, and corresponding equivalence classes $[S] = \{S' \mid S' \sim S\}$. Then, we have that, if S is a scene, $[S]$ is a physical scene. Note that, if we had defined relations *not among scenes* based on the images they hallucinate, *but among measured images* as done in classical objectivist “Marrian” style of Computer Vision, we would not have equivalence classes, because sets of images from the same scene are related by co-visibility, which is not a transitive relation: The fact that the frustra (pre-images) of two images y_1, y_2 intersect, and so do the frustra of y_2, y_3 , does not imply that that frustra of y_1 and y_3 intersect.

The previous remark can be summarized in the following claim:

Claim 2.12 (Physical Scenes as Equivalence Classes). if S is a scene, then $[S] = \{S' \mid p(\cdot|S', c) = p(\cdot|S, c) \forall c\}$ is the corresponding physical scene.

Note that changing the measurement function may change the corresponding physical scene, much as laws of physics may need to be amended if new instruments produce evidence that invalidates the laws.

Remark 2.13 (Presentations as “controlled hallucinations”). Note that scenes induce equivalence classes of *infinite collections* of images, not of *different finite collections of given images of a scene*. Rather, they are the *distributions of images that two scenes could “hallucinate”* using their presentation function. For this reason, even if $p(\cdot|S, c)$ defines the equivalence class that represents the scene, we refer to it as “presentation” rather “representation,” following Koenderink [15]. A “representation” presumes that something is present to begin with, which can be manipulated and thence re-presented. But, unlike the data, the scene S is an abstract entity which is not accessible to be manipulated or re-presented. Nonetheless, it can be used to *present* images that are compatible with the given ones, a process referred to as “controlled hallucinations.” The input c provides the control, and the presentation function is the abstract mechanism that the scene uses to generate images, or more properly hallucinate them since the process is an abstraction of data formation, not an actual measurement.

The interpretation in the previous remark may provide hints to the reader of what characteristics a neural network should possess to be a Foundation Model of physical scenes, which we develop in the next section, after a few remarks.

Remark 2.14 (“Large Word Models” vs. Large Language Models). One could call a map from images to scenes $\mathbf{y} \mapsto [S]$ a “large⁷ world model” (LWM), much in the same way in which a large language model (LLM) is a map from sentences to meanings [37]. Just as the physical scene is an equivalence class of images that can be inferred by a LWM, which can then be used to hallucinate infinitely many images, a “meaning” in its most elementary and naive sense, can be defined an equivalence class of sentences that can be inferred by a LLM, which can then be used to hallucinate infinitely many sentences controlled by the given “prompt” [37]. Since both sentences and images can be embedded (see Remark 2.2) as vectors in a metric space by their corresponding models, then the goal of a Foundation Model, whether operating on image or language or other data, is to give meaning to the data, which it can do by mapping data to a metric space and constructing equivalence classes therein, as well as to model the “prior” distribution $p([S])$ of realistic scenes and the generative distribution $p(\mathbf{y}|[S])$. For sensory data, the meaning rests in the scene that generates them, and is provided by *grounding*; for language data, the meaning can only be defined by using human-provided annotations, a challenge discussed next.

Remark 2.15 (Grounding). Just like a trained LLM attributes meaning to sentences by constructing equivalence classes of them, so a LWM attributes meaning to images, in the sense above, by constructing equivalence classes of scenes. In general, meaning is not *intrinsic* in data, but rather *attributed* to data, a

⁷Although we have given no indication that such a model should be “large,” the functional anatomy of primate neocortex [11] suggests that a large portion of real estate in human brains is devoted to processing visual information, so if language models are large, so should be visual models.

process that requires an external entity [27]. This entity could be the one that *originates* the measured data, for instance the surrounding environment, informally referred to as the “true” or “real” scene, for the case of sensory data. Relating measured data to the source is called *grounding*. But unlike natural language data, which originates in the human brain that is not accessible for experimentation, the surrounding environment can be probed with independent sensors by an embodied entity. As pointed out in [37], environmental grounding is not veridical but it is falsifiable: We cannot verify the existence of an object, say “chair,” in an image, for there are no chairs in the images, just pixels, but an embodied agent can attribute meaning to a chair by testing its ability to sit on it (affordance). On the other hand, grounding in language *must* rely on induction – which is not falsifiable – based on human annotations – which are subjective and therefore not veridical despite the language being a closed model: If training data have inconsistent human annotations, the “ground truth” used to build the discriminant that defines meanings is inconsistent, and so is the resulting system of meanings. Not so for the surrounding environment, so long as it exists and can be probed with independent modalities.

This fundamental difference between LLMs and LWMs, which is the possibility of cross-modal grounding, is highlighted but not exploited in this paper, which only considers world models that can be inferred from passively gathered, remote, non-contact, distributed sensory data, in particular images.

Remark 2.16 (Self-supervision, Contrastive Learning, and the universality of the prediction task). In order to be viable, a scene must generate images through its presentation function that, while generally different from the ones that the “true scene” would generate if we could measure them, are *indistinguishable* (distributionally equivalent, see Remark 2.2.) For images that we *did* measure, this defining characteristic of scenes indicates a natural learning criterion, which is simply *data prediction*, or “masked” reconstruction. That is, we use the presentation function to hallucinate images from a vantage point we *did* observe, and then compare the hallucinated images with the ones we *actually* measured. The masking can be causal and delayed, where prediction of future images is compared with actual images to be taken a few steps in the future, once they are observed. Prediction is universal in the sense that, if a model is capable of predicting future images, it is therefore sufficient to support any function of future images instantiated at inference time. The question, then, is whether it is minimal. In the absence of any information about the task, the only representation that is sufficient is the data itself or any lossless representation of it, for the downstream task may be exact retrieval. It is common to refer to so-called “self-supervised learning,” including contrastive learning, as “task agnostic.” This is incorrect since the task is implicit in the choice of nuisance variability that defines the surrogate losses, data augmentations, or transformations that are manually designed. For example, contrastive learning that imposes that rotated versions of the same image represent the same equivalence class cannot, trivially, be useful for the task of determining image rotation. Similarly, masked autoencoding is a way of simulating occlusion, which is useful if there are occlusion nuisances, but not if the sensory modality is X-ray tomography or magnetic resonance where there are no occlusions. On the other hand, all variability in visual data is manifest over time, for an embodied agent, which makes prediction a natural choice.

3 Towards Foundational Models of Physical Scenes

The previous section established the scene as an abstract concept with different embodiments, which could be a computer program or, say, a Deep Neural Network (DNN). This raises hopes that NeRFs may be a representation of the scene and therefore a viable tool in developing generic models of the world from images for the purpose of any downstream task. In other words, we ask whether NeRFs and/or Diffusion Models could be general Foundational Models for physical scenes, in the sense of implementing a *presentation function*.

3.1 NeRFs as (re)presentations of scenes?

We will describe NeRFs in more detail in Sec.4, but for now it suffices to say that they are maps h learned from images of a scene S with knowledge of pose c that, at inference time, can be used to map a novel pose

c' onto an image $y = h(S, c')$. But do they implement a presentation function? The fact that, in a NeRF, the map h is implemented as a feed-forward memoryless operator using a multi-layer perceptron (MLP), is sufficient to tamper expectations.

Proposition 3.1. NeRFs cannot represent scenes.

The proof follows from Proposition 2.9 in [1]. By Remark 2.2 a representation of the scene needs to be able to hallucinate new *realistic* measurements given a set of past measurements. On the surface, NeRFs can use past data to create an occupancy model, and then use this model to generate images from novel viewpoints. However, those images need not be realistic. This is obvious in practice and illustrated in Fig. 4. A NeRF of an object captured rotating around it at a certain distance can be used to generate images from, say, twice the distance but any attempt to “zoom in” will generate artifacts that reveal the NeRF \tilde{S} as not *the real (presentation function of the) scene*. In other words, the attempt to extrapolate is *unrealistic*. The meaning of *unrealistic* here is not some vague notion of perceptual similarity having to do with the human visual system, but the precise conditions of Theorem 2.10: Images captured by the camera are obviously distinguishable from those generated by the NeRF, so $\exists c \mid p(\cdot|\text{GoPro}, c) \neq p(\cdot|\text{NeRF}, c)$, violating the conditions of the theorem.

At the core of the problem is the fact that while NeRFs provide a measurement model $p(y|S)$, they do not provide a distribution $P(S)$ over the scenes, which encodes “realism.” That is, wherever the given data provides constraints for interpolation, the model should faithfully reproduce it, which it does since it is trained to do so explicitly through the reconstruction error. However, where the data do not provide sufficient constraints, the model should extrapolate so that, if data were to become available at different granularity, it could be compared even if it was not trained on it. Comparison is in a distributional sense, so a viewer or model would not be able to tell which image comes from the “true scene,” whatever that is, and which comes from the NeRF. For a trained model, this rests on induction, which is why we need to train a model on *different scenes*, and then use the result to hallucinate details where data from the extant scene is insufficient. In doing so, however, the model must remain faithful to the conditions of Theorem 2.10. To do so, we bring in Diffusion Models.

3.2 Pushing a NeRF towards the physical scene with Diffusion Models

We will use a Diffusion Model (DM) as a map from the output of a NeRF, $\tilde{y} = h(S, c)$ to an image y , which is in the form of a conditional distribution $p(y|h(S, c))$ from which images can be sampled. We defer the details to Sec. 4, but for now what matters is the fact that DMs are trained to approximate the distribution of natural images, conditioned on a “prompt,” and given the scale-invariance properties of low-level statistics [56], they can be used to hallucinate details where absent, with the hallucination driven by the inductive bias of the trained model. Indeed, DMs have been used to “denoise” or “super-resolve” images, which are ill-posed inverse problems where the inductive bias is used to hallucinate details and structure absent in the data. Clearly a standard DM itself is not a viable presentation function. However, we hypothesize that, when composed with (conditioned on) the output of a NeRF, they are viable presentation functions.

We call the composition of the NeRF h and the diffusion model *NeRF Diffusion*, and we indicate the corresponding measurement function as f_S , which is a stochastic function. The following proposition shows that NeRF Diffusion models *are* a viable presentation function, and therefore equivalent to *the physical scene* given the images used for training.

Proposition 3.2 (NeRF Diffusions as (re)presentations of the physical scene). Given a set of observations $\mathbf{y}_{\leq t} = (\mathbf{y}_1, \dots, \mathbf{y}_t)$, a NeRF Diffusion defines a stochastic $f_{\mathbf{y}_{\leq t}} : \mathcal{C} \rightarrow \mathbb{R}^D$ such that $f_{\mathbf{y}_{\leq t}} \sim \int p(y|S, c)p(S|\mathbf{y}_t)dS$. This is a representation of the scene where we take the scene S to be $S = (\mathbf{y}_1, \dots, \mathbf{y}_n)$.

Proof. We need to show that $f_{\mathbf{y}_{\leq t}}$ is enough to compute the distribution $P(S) = P(\mathbf{y}_1, \dots, \mathbf{y}_n)$. To this end, note that

$$P(\mathbf{y}_1, \dots, \mathbf{y}_n) = \prod_{t=1}^n P(\mathbf{y}_t | \mathbf{y}_{<t}) = \prod_{t=1}^n f_{\mathbf{y}_{\leq t}}(\mathbf{y}_t).$$

□

f_S is also minimal as a function, although any particular implementation may be super-minimal due to implementational inefficiencies. In the next section, we describe an implementation, which we consider a first step towards a proper model of the physical scene, exploiting methods from the current state of the art.

4 NeRF Diffusion implementation

In this section we describe a method to adapt a Diffusion Model to complement a NeRF so as to implement a presentation function that satisfies the conditions of Theorem 2.10.

Neural Radiance Fields (NeRFs) [24] are multi-layer perceptrons trained to approximate the radiance field, *a.k.a.* Plenoptic Function. To render a pixel, points are sampled along the ray passing through the pixel and originated from the camera with configuration c . Density and view-dependent radiance values are obtained by querying the MLP with the points' spatial coordinates and viewing direction. Volumetric rendering is then performed on the density and radiance values to produce the color at the pixel. NeRF approaches [24, 2, 3, 42, 22] are mostly trained with calibrated images (c is known during training as part of the camera intrinsic and extrinsic calibration) supervised by pixel-wise discrepancy. Camera pose and calibration could also be inferred from the images as part of a pre-processing stage, for instance using standard Structure From Motion tools.

Diffusion Models (DMs) are statistical models of the distribution of images obtained by learning the reverse diffusion operator of a process that maps training images to noise [26, 19, 19]. DMs can be conditioned on a number of inputs, from text to sketches, low-resolution images or, in our case, unrealistic NeRF renderings. In other words, we condition the DMs on images produced by a NeRF, so it learns to tilt the generative distribution implemented by a NeRF, which is unrealistic, to one that is realistic, hence compatible with the physical scene.

To realize a *NeRF Diffusion*, we need to train a DM to denoise views synthesized by a NeRF and extrapolate details not manifest therein. The outcome is improved phenomenological quality (realism) of the rendering, achieved by learning NeRF priors (Fig. 5). We train a DM to learn the distribution of NeRF rendering artifacts and map the resulting aliased renderings to the corresponding calibrated images. To that end, we use Nerfacto within Nerfstudio [39]: We turn off pose refinement and train Nerfacto for 40K iterations in about 30 minutes on the ObjectScans dataset described in Sec. 5, using a single NVIDIA V100. We then train a DM based on Stable Diffusion [31], a large-scale text-to-image latent diffusion model. To leverage the benefit of text-to-image generation capabilities [44], which provides a strong prior to hallucinate semantically plausible details, we initialize the weights of our model with a pretrained Stable Diffusion checkpoint. To condition the DM, we directly add the NeRF rendering's latent features into Stable Diffusion's U-net structure, as in [53].

Given a latent image y_0 and its corresponding NeRF rendering's latent feature S , diffusion algorithms progressively add noise to the image to produce a noisy version y_t . Here t is the number of steps in which noise is added. Since we are learning the NeRF priors to model the artifacts, we set the text prompt u_t of Stable Diffusion to generic prompt (i.e. "high resolution and high quality"). The overall objective is

$$\mathcal{L} = \mathbb{E}_{y_0, t, u_t, S, \epsilon \sim \mathcal{N}(0, 1)} [\|\epsilon - \epsilon_\theta(y_t, t, u_t, S)\|_2^2].$$

We train our DM on all 121 sequences from ObjectScans' NeRF for 10K iterations in 6 hours using 8 NVIDIA V100 GPUs. The final trained model captures the NeRF's structural priors and sampling operators so that, when a new condition (rendering) is presented, the model extrapolates realistic images, thus realizing a simple embodiment of a NeRF Diffusion.

5 Experimental evaluation

5.1 Details on the datasets

ObjectScans consists of 121 video sequences of common household objects such as cookware, containers, and toys. The videos were captured by the authors using GoPro Hero9 action cameras at 30 FPS and 1920×1080 resolution. The length of each video ranges from 20 seconds to about 1 minute. Unlike existing datasets that are either captured by stationary DSLR cameras [3, 23] or synthesized [24], recording videos makes it possible to scale the data acquisition process for training NeRFs on real-world scenes. Yet, video data pose a challenge to the established NeRF techniques on account of motion blur and compression artifacts not present in existing datasets. To minimize motion blur, we move the camera slowly and steadily around the object during data acquisition. We also observed that NeRFs work best when the camera tightly samples the space of viewpoints. We employed these best practices in our data collection. For each video sequence, we uniformly sample about 300 frames to conduct our experiments totalling about 36K frames in the whole dataset. We split the sampled frames into train/test subsets with a 9 : 1 ratio, and then use COLMAP [36] – an open-source Structure-from-Motion software – to compute camera intrinsics, poses and sparse point clouds from the sampled video frames.

Mip-NeRF 360 was created by [3] and consists of 9 scenes (5 outdoors and 4 indoors), of which only 7 (namely, `bicycle`, `bonsai`, `counter`, `garden`, `kitchen`, `room`, `stump`) are publicly available. Each of the scenes contains a central object or area that has relatively complex geometry and/or texture and a background that contains fine-grained details (*e.g.*, foliage). The dataset also comes with camera pose pre-computed using COLMAP. We use these pre-computed camera poses in our experiments.

5.2 Evaluation metrics

We evaluate our method on *faithfulness* through similarity metrics and *realism* through distribution metrics.

Similarity metrics: PSNR, SSIM [47], LPIPS [54]. PSNR calculates the ratio of maximum pixel value to noise. PSNR is a sensible performance measure for compression and transmission, not for representation. SSIM is sensible for perceptual similarities based on luminance, contrast, and structure, but our goal is not to achieve indistinguishability as defined by human perception, but by presentation functions. LPIPS is a criterion based on convolutional features trained for image classification [16]. As such, it is generally more appropriate to evaluate scene representations.

Distribution metrics: IS [35], FID [12], KID [4]. In FID, a score is calculated by computing the Fréchet distance between two Gaussians fit to features computed by the Inception network. KID measures the dissimilarity between the distributions of real and generated samples without assuming any parametric form, and is also more sample efficient. The inception score is calculated by first using a pre-trained Inception v3 model to predict the class probabilities for each generated image. The Inception score is computed only on generated images without the need for real images. For all distributional metrics, we precompute real statistics using 10k patches of real image samples from each test sequence of our dataset.

5.3 Quantitative results

Table 1 provides the comparison of our approach with Nerfacto [39] – an efficient and top-performing NeRF model – and Real-ESRGAN [45] – a state-of-the-art (SOTA) GAN-based image restoration model – on ObjectScans. We fine-tune both ours and Real-ESRGAN on Nerfacto renderings. Ours has a lower FID, KID score than both Nerfacto and Real-ESRGAN, reflecting the higher plausibility of the NeRF rendering. We also observe that our approach is competitive with baselines in similarity metrics (*e.g.*, PSNR and LPIPS). Even though these metrics over-penalize high-frequency details (which is commonly observed in diffusion-based image generation), our approach preserves the original details faithfully while removing the artifacts and enhancing realism. Since SSIM evaluates pixel value changes and diffusion-based approaches are designed to hallucinate, our approach under-performs when compared to Nerfacto baseline.

Table 1: *Quantitative results on ObjectScans*. Ours outperforms the baselines in various distribution metrics (FID, KID, IS) showcasing our model’s capability to predict visually plausible images. Furthermore, our model respects measurements (ground-truth images) sampled from the scene registering competitive similarity scores (PSNR, SSIM, LPIPS).

Sequences	Nerfacto					Real-ESRGAN					Ours							
	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	IS ↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	IS ↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	IS ↑
<i>RiceCooker</i>	23.78	0.86	0.32	100.89	0.057	6.74	22.91	0.85	0.26	45.84	0.020	7.08	24.97	0.86	0.17	21.15	0.0060	7.65
<i>WaterFilter</i>	24.17	0.87	0.26	81.34	0.0398	5.61	23.70	0.86	0.22	45.69	0.019	6.35	26.17	0.85	0.19	23.96	0.0058	6.49
<i>InstantPot</i>	24.11	0.85	0.21	88.75	0.046	6.24	22.93	0.81	0.17	31.99	0.012	6.69	25.88	0.83	0.11	18.15	0.007	7.05
<i>Jar</i>	23.19	0.84	0.28	76.78	0.034	6.25	22.34	0.81	0.21	45.70	0.020	6.87	25.38	0.83	0.15	27.25	0.01	6.56
<i>ThermGun</i>	24.13	0.87	0.26	74.15	0.035	5.32	23.69	0.86	0.22	42.17	0.019	5.77	27.5	0.87	0.15	19.16	0.0049	6.18
<i>Robot</i>	24.29	0.85	0.26	68.51	0.029	6.13	23.56	0.83	0.20	44.08	0.020	6.47	27.80	0.84	0.13	23.17	0.008	6.54
<i>GlassPotLid</i>	23.59	0.72	0.37	101.59	0.06	4.15	22.93	0.69	0.38	68.50	0.032	4.89	24.04	0.66	0.27	31.74	0.012	4.76
<i>LanternOff</i>	23.03	0.79	0.26	83.89	0.043	6.52	21.94	0.75	0.25	43.95	0.014	7.32	24.12	0.78	0.16	24.32	0.01	7.29
<i>LanternOn</i>	22.67	0.78	0.27	83.22	0.045	6.69	21.82	0.75	0.25	43.63	0.016	7.71	24.05	0.78	0.16	23.42	0.0088	7.20
<i>LegoBus</i>	24.14	0.78	0.26	49.78	0.021	6.74	22.47	0.73	0.23	26.81	0.006	6.58	24.33	0.74	0.14	17.05	0.0061	6.33
<i>MacBook</i>	21.50	0.80	0.28	72.38	0.038	5.2	21.04	0.78	0.25	37.34	0.015	5.88	24.62	0.80	0.15	23.21	0.0092	6.10
<i>SquareGlassJar</i>	23.68	0.80	0.25	83.19	0.043	6.16	22.70	0.74	0.25	54.26	0.019	7.39	24.82	0.78	0.15	25.89	0.0102	7.52
<i>StainlessSteelHotpot</i>	21.77	0.72	0.38	95.54	0.053	4.91	21.35	0.71	0.37	57.23	0.022	5.83	24.08	0.69	0.28	25.92	0.0078	5.27
<i>WaterBoiler</i>	22.59	0.79	0.26	86.14	0.034	7.92	21.57	0.76	0.25	47.79	0.016	8.84	23.62	0.78	0.16	23.54	0.0079	7.80
Average	23.33	0.81	0.28	81.87	0.041	6.04	22.50	0.78	0.25	45.36	0.018	6.69	25.10	0.79	0.17	23.42	0.0081	6.62

Table 2 compares ours against Nerfacto baseline on public Mip-NeRF 360 dataset. Our approach outperforms Nerfacto in distribution metric – similar to our observations in Table 1. However, the Nerfacto’s similarity metric performs better especially in PSNR and SSIM. The performance gap in similarity metric could be due to fact that we fine-tune the DM which was already trained on ObjectScans and the new fine-tuned model struggles to generalize. To evaluate the generalization hypothesis, we trained the DM only on Mip-NeRF 360 sequences and evaluated the performance on *counter* sequence. We observe performance improvement in our approach (PSNR = 23.79, SSIM = 0.70, LPIPS = 0.25) when compared to baseline (PSNR = 22.90, SSIM = 0.82, LPIPS = 0.37) in terms of similarity metrics (particularly PSNR and LPIPS). We leave the out-of-distribution generalization of our approach for the future work.

Table 2: *Quantitative results on Mip-NeRF 360 dataset*.

Dataset	Nerfacto					Ours						
	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	IS ↑	PSNR↑	SSIM↑	LPIPS↓	FID↓	KID↓	IS ↑
<i>bicycle</i>	23.15	0.69	0.26	48.10	0.0084	4.21	21.24	0.47	0.21	26.17	0.0027	4.03
<i>bonsai</i>	29.77	0.92	0.07	31.21	0.0154	3.98	24.02	0.75	0.10	20.25	0.0066	4.19
<i>counter</i>	22.90	0.82	0.37	102.72	0.048	7.28	20.87	0.64	0.30	23.50	0.012	6.96
<i>garden</i>	24.62	0.70	0.38	111.56	0.0634	4.89	22.11	0.64	0.24	23.23	0.0074	4.34
<i>kitchen</i>	28.17	0.83	0.18	55.51	0.026	5.38	23.74	0.65	0.16	20.93	0.0057	5.21
<i>room</i>	30.84	0.87	0.27	80.38	0.0312	5.36	25.08	0.69	0.24	42.72	0.01	7.38
<i>stump</i>	25.95	0.71	0.24	65.51	0.055	2.64	22.94	0.50	0.26	25.04	0.011	3.00
Average	26.48	0.79	0.25	70.72	0.0353	4.81	22.85	0.62	0.22	25.97	0.0078	5.01

5.4 Qualitative results

Fig. 1 presents qualitative results on ObjectScans. NeRFs to model fine-grained details and glossy surfaces which can be seen in column two. A generic image denoising approach such as Real-ESRGAN fine-tuned on our dataset is able to remove some artifacts but fails to extrapolate structures beyond the Nyquist frequency, resulting in over-smoothing. The proposed approach removes artifacts without oversmoothing. *Furthermore, the proposed approach is able to hallucinate realistic details which may be absent in the scene but nevertheless compatible with the given data and with the natural image statistics*, for example, the reflection on the rice cooker (top rows) and highlights at the bottom of the water filter (bottom rows).

Fig. 2 presents on qualitative results Mip-NeRF 360. The scenes in Mip-NeRF are out-of-distribution relative to ObjectScans, and thus the DM pre-trained on the latter is biased to generates samples with inconsistent color distribution. Yet, the generated samples are realistic, which is the goal of the DM, and compatible with the data, which is the goal of the NeRF. By fine-tuning the DM on Mip-NeRF, we are able to match the

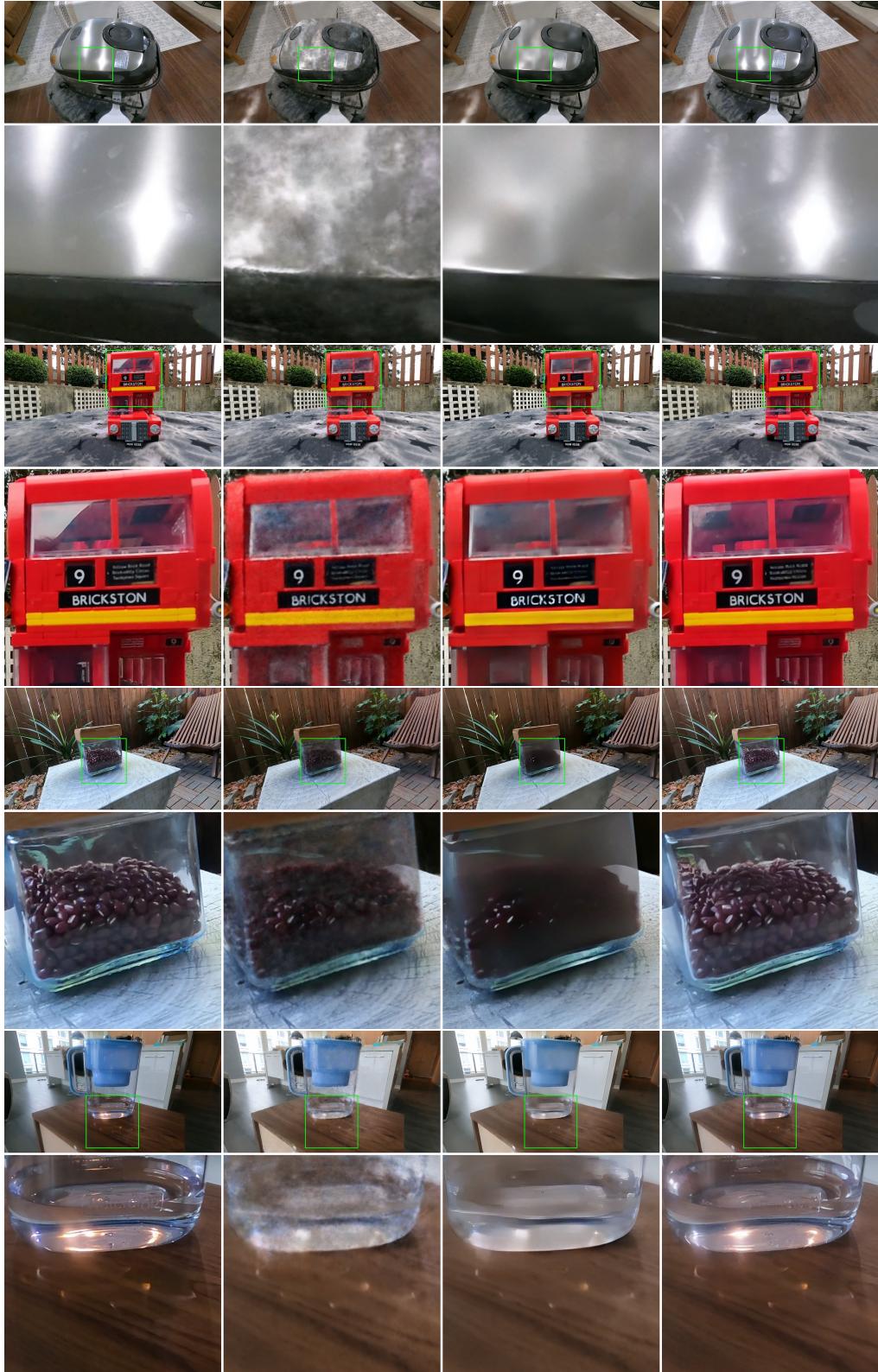


Figure 1: *Visual comparison* of the proposed approach against baselines on our datasets. **Left to right:** (a) ground-truth reference image, (b) NeRF rendering, (c) NeRF rendering augmented by Real-ESRGAN [45] – SOTA in image restoration, and (d) NeRF rendering augmented by our method. **Top to bottom:** 4 samples with corresponding zoom-in views. Note, the goal is *not* to make the output exactly the same as the ground truth, but to generate a plausible image of the underlying scene given an imperfect sample (column two).

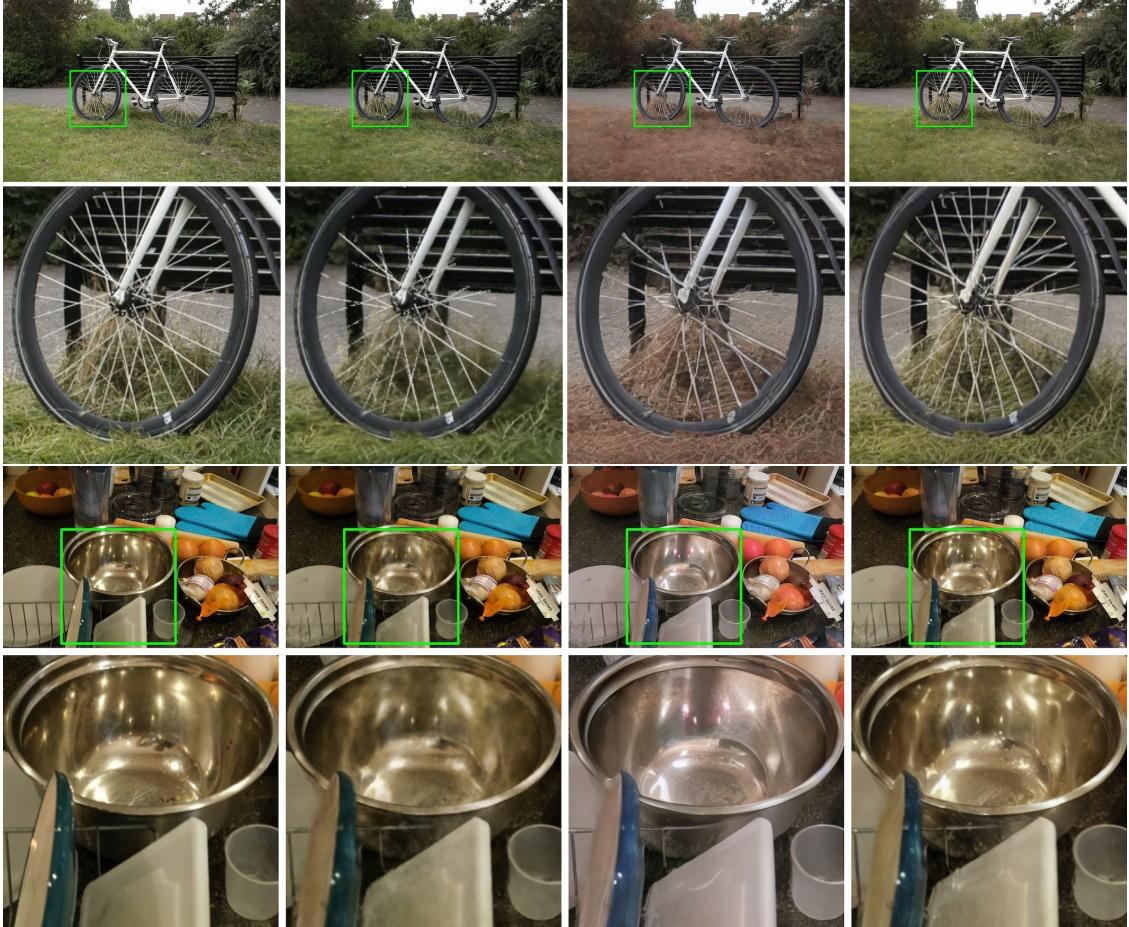


Figure 2: *Qualitative results* of the proposed approach on the public Mip-NeRF 360 dataset. **Left to right:** (a) ground-truth reference image, (b) NeRF rendering, (c) rendering augmented by diffusion model pre-trained on our ObjectScans dataset, and (d) rendering augmented by the diffusion model in (c) which is further fine-tuned on Mip-NeRF 360 dataset. Bottom row shows each sample's corresponding zoom-in view.

color distribution, and also fill in fine-grained details missing in NeRF’s renderings (column two). Note the added details such as reflections on the bowl, not necessarily present in the scene but nevertheless realistic.

5.5 Super-resolution

Fig. 4 illustrates the extrapolation characteristics of our NeRF Diffusions. For a given test sample with ground-truth image and camera pose available, we first obtain the NeRF rendering and its corresponding NeRF Diffusion output. Also, we sample additional camera poses near the test pose closer to the object to obtain the “zoomed-in” version of the NeRF rendering. In the case of Fig. 4, we focus on rendering the steel vessel and oil bottle. Since NeRFs fail to interpolate at finer resolution, the NeRF’s rendering would have more artifacts. Through feeding this aliased NeRF rendering to a DM, we obtain visually plausible and enhanced scenes through DM’s hallucinations/extrapolation.

5.6 Limitations

It is known that diffusion models require considerable amounts of training data. As can be seen in Fig. 2 column 3, the sample image generated by our model is biased towards the data distribution that the model has been trained on, since our models are trained on small data. Fine-tuning the model on a specific data domain or training the diffusion model on larger dataset can mitigate the problem.

Another problem that challenges diffusion models is their inability to model fine-grained structure like text, which we also observed in (Fig. 3). Other limitations include sampling speed and high computation costs. While “hallucination” is a problem in synthetic data generation, in our case it is the goal, so long as the NeRF Diffusion remains faithful to the data where available. The synthesis is tasked to add details that are compatible with natural image statistics, without affecting the rendering where information is manifest in the given data.

Note that, in this aspect, world models are fundamentally different from language models, due to *scale variability in the data* which is absent in language. Given a sentence, there are no words revealed as we change the measurement conditions. Given an image, on the other hand, there are always more details about the scene to be revealed if we move closer to objects. These details are not inferrable from the given images by interpolation, which is why a Foundation Model for visual scenes must incorporate a recursion mechanism or other mechanism to go to the limit.

6 Related work

Diffusion models have significantly advanced image generation offering a scalable and robust training paradigm in either pixel [29, 34] or latent space [32], and taking image generation to a new level of realism and diversity. Interest has grown in extending diffusion models to 3D content creation [43, 51, 28, 55, 18, 21, 13], usually by guiding NeRF training with natural image priors from 2D diffusion models. DiffusioNeRF [50] aims to use a pre-trained denoising diffusion model to regularize NeRF trained on only a few images. Nerfbusters [48] attempts to use diffusion models to remove the the ghosting effects (*a.k.a.* floaters) from NeRF rendering.

Designing a viable scene representation, or creating a map $y = h(S, c)$ for image prediction has been pursued in different contexts or research fields. Some focus on augmenting existing NeRF models by introducing additional complexity to the map h to (hopefully) circumvent known limitations of the vanilla model such as aliasing or non-Lambertian effects [24, 2, 3, 42, 22]. Some focus on directly learning an end-to-end denoising or super-resolution network to recover the local statistics of generated images [41, 46]. Those works validate our proposition that a scene is an abstract concept specified via a computational procedure, which has no objective grounding.

Specifically on NeRFs, [24] and related work [2, 3, 42, 22] are state-of-the-art approaches for novel view synthesis. NeRF renders images by encoding density and view-dependent radiance based on multi-layer



Figure 3: *A known challenge* for diffusion models is fine-scale generation, for instance text. **Left to right:** (a) ground-truth image, (b) NeRF rendering, (c) NeRF rendering augmented by our approach. The bottom row shows zoomed-in views of the top row. Though our approach fills in some high-frequency signals that NeRF failed to capture, it is not capable of hallucinating meaningful text that can be seen in the ground-truth image.

perceptron (MLP). However, there are several drawbacks. First, training with varying scene resolution leads to aliasing because any 3D scene point is infinitesimally small regardless of its distance to the camera. Mip-NeRF [2] attempts to solve this problem by modeling the 3D scene point using a frustum. The second challenge is with modeling non-Lambertian scenes that include reflection and refraction. Ref-NeRF [42] addresses reflection modeling by extending the NeRF MLP to predict surface normals and reflection properties (e.g., roughness) in order to calculate reflection direction for specular albedo estimation; this is added to the diffuse color for final rendering. Similarly, NeRFRen addresses the reflection problem by separating transmitted and reflected color. To the best of our knowledge, there is little work on refraction modeling within the volumetric rendering literature. Refraction modeling is addressed based on multi-plane image [49] or light field [38]. The next drawback of NeRFs is on their computationally intensive volumetric rendering process. Since we have to query the MLP for every sample point along every camera ray, rendering a single image takes more than several seconds, making it impossible for interactive real-time rendering. Instant-NGP [25] drastically accelerates NeRF rendering by encoding spatial information using hash maps, which reduce the MLP size and enable faster rendering. Nerfstudio [39] is an open-source implementation for various NeRF approaches, including Instant-NGP. We have validated that Nerfstudio yields state-of-the-art accuracy within reasonable training time (within half an hour using Tesla V100) and thus used Nerfstudio through out our experiments.

As for datasets, popular choices used in the NeRF literature are either synthetic (*e.g.*, NeRF-Synthetic [24] and BlendedMVS [52]) or only contain a handful of real-world scenes (*e.g.*, Mip-NeRF 360 [3], LLFF [23]). More recent datasets such as MobileBrick [17] and ScanNeRF [8] contain a few dozen sequences in a laboratory setting: The former captures RGB-D data of LEGO models with known ground-truth CAD models and the latter uses specialized hardware to capture small objects limiting their applicability in real-world scenarios. ScanNet [7] contains thousands of RGB-D scans of rooms, but its applicability to train NeRFs is yet to be explored. CO3D [30] contains about 19K crowd-sourced videos, but focuses on object reconstruction task and as such often has simplistic background. To fill the gap left by existing public datasets, we collected a relatively large dataset using widely available commodity hardware, and conducted some of our experiments using this dataset.

7 Discussion

As our title suggests, we consider ours only a first step towards building Foundational Models of physical scenes. Nonetheless, to the best of our knowledge, this paper is the first to propose a practical method to represent physical scenes that goes beyond naive realism, dominant since Marr [20], embracing the analytical approach advocated by Koenderink [15], formalized and implemented with contemporary Deep Learning tools.

Evaluating methods that generate details that are not manifest in the given data makes assessment challenging, as all existing quantitative benchmarks fall into the objectivity trap, by effectively *defining* an objective reality where there is none. At the same time, perceptual similarity metrics, which are biased towards the characteristics of the human visual system, are still quite rudimentary, often measuring limited variability (e.g. SSIM) or other coarse distributional statistics (FID, KID, IS). Considerably more work needs to be conducted to develop methods that assess the realism of generated images, as opposed to their fit to “reality” defined by an arbitrary benchmark design. We note that our analysis is limited to vision, and must be extended to other modalities, and in particular for active perception by probing the environment with structured signals or contact sensors. We leave these extensions to future work.

References

- [1] A. Achille and S. Soatto. On the learnability of physical concepts: Can a neural network understand what’s real? *arXiv preprint arXiv:2207.12186*, 2022.
- [2] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.
- [3] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.
- [4] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans, 2021.
- [5] G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021.
- [6] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.
- [7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [8] L. De Luigi, D. Bolognini, F. Domeniconi, D. De Gregorio, M. Poggi, and L. Di Stefano. Scannerf: a scalable benchmark for neural radiance fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 816–825, 2023.
- [9] P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.
- [10] A. Einstein. *The philosophy of Bertrand Russell, Part II: Descriptive and Critical Essays on the Philosophy of Bertrand Russell*. 1946.
- [11] D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- [12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

- [13] A. Karnewar, A. Vedaldi, D. Novotny, and N. Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. *arXiv preprint arXiv:2303.16509*, 2023.
- [14] R. W. Keener. *Theoretical statistics: Topics for a core course*. Springer, 2010.
- [15] J. J. Koenderink. Vision and information. *Perception beyond inference: The information content of visual processes*, pages 27–58, 2011.
- [16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [17] K. Li, J.-W. Bian, R. Castle, P. H. Torr, and V. A. Prisacariu. Mobilebrick: Building lego for 3d reconstruction on mobile devices. *arXiv preprint arXiv:2303.01932*, 2023.
- [18] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.
- [19] A. Lindquist and G. Picci. On the stochastic realization problem. *SIAM Journal on Control and Optimization*, 17(3):365–389, 1979.
- [20] D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 1980.
- [21] L. Melas-Kyriazi, C. Rupprecht, I. Laina, and A. Vedaldi. Realfusion: 360 $\{\backslash\deg\}$ reconstruction of any object from a single image. *arXiv preprint arXiv:2302.10663*, 2023.
- [22] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022.
- [23] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.
- [24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [25] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.
- [26] E. Nelson. *Dynamical theories of Brownian motion*. Princeton university press, 1967.
- [27] P. M. Pietroski. *Conjoining meanings: Semantics without truth values*. Oxford University Press, 2018.
- [28] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [29] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [30] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.
- [31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

- [33] B. Russell. *The Problems of Philosophy, Chapter 1: Appearance and Reality*. OUP Oxford, 2001.
- [34] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans, 2016.
- [36] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [37] S. Soatto, P. Tabuada, P. Chaudhari, and T. Y. Liu. Taming ai bots: Controllability of neural states in large language models. *ArXiv*, 2023.
- [38] M. Suhail, C. Esteves, L. Sigal, and A. Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022.
- [39] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa. Nerfstudio: A modular framework for neural radiance field development. *arXiv preprint arXiv:2302.04264*, 2023.
- [40] M. Trager, P. Perera, L. Zancato, A. Achille, P. Bhatia, B. Xiang, and S. Soatto. Linear spaces of meanings: the compositional language of vlms. *arXiv preprint arXiv:2302.14383*, 2023.
- [41] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.
- [42] D. Verbin, P. Hedman, B. Mildenhall, T. Zickler, J. T. Barron, and P. P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.
- [43] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. *arXiv preprint arXiv:2212.00774*, 2022.
- [44] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022.
- [45] X. Wang, L. Xie, C. Dong, and Y. Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021.
- [46] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.
- [47] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [48] F. Warburg, E. Weber, M. Tancik, A. Holynski, and A. Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. *arXiv preprint arXiv:2304.10532*, 2023.
- [49] S. Wizadwongsa, P. Phongthawee, J. Yenphraphai, and S. Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021.
- [50] J. Wynn and D. Turmukhambetov. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. *arXiv preprint arXiv:2302.12231*, 2023.

- [51] D. Xu, Y. Jiang, P. Wang, Z. Fan, Y. Wang, and Z. Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 $\{\backslash\deg\}$ views. *arXiv preprint arXiv:2211.16431*, 2022.
- [52] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.
- [53] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.
- [54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [55] Z. Zhou and S. Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. *arXiv preprint arXiv:2212.00792*, 2022.
- [56] D. Zoran and Y. Weiss. Scale invariance and noise in natural images. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2209–2216. IEEE, 2009.



Figure 4: NeRFs cannot interpolate at finer resolution without visible artifacts, that betray the NeRF as not equivalent to other scenes, in the sense of Theorem 2.10, and therefore non viable as a representation of the physical scene. **Left to right:** (a) ground-truth reference image, (b) NeRF rendering, (c) NeRF Diffusion output. The top row corresponds to test samples of the `counter` sequence in Mip-NeRF 360 dataset. In the second and third rows, we randomly sample zoomed-in camera poses around the object and obtain corresponding NeRF rendering (column b) and NeRF Diffusion output (column c).

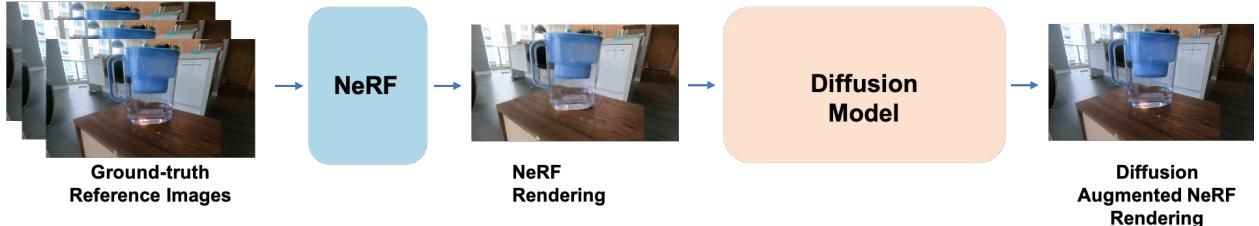


Figure 5: A block diagram of our simple model.

A Q&A

In this section, we discuss some of the most controversial elements of our paper.

- **Q: What does this paper have to do with physical scenes? There is no talk of surfaces, reflectance, motion, any of the aspects described as critical for interaction with physical space.**
- **A:** Whatever properties we seek to infer about the scene, including geometry, photometry, dynamics, etc., can only be inferred if they are manifest in the data. Any representation that can be inferred from the data and is *maximally informative* (sufficient, per Definition 2.1), can be used to *infer any properties or attributes of the scene, even if they are not explicit in the representation*.
- **Q: But there is no way to pinpoint 3D shape in the parameters of a NeRF or a Diffusion Model. Where is the shape?**
- **A:** Shape information (not the shape itself) is implicit in the parameters of the trained model, and in the activations after feeding all the suitable conditioning inputs. The real question to ask, here, is *where is the shape in the “real scene”?* Describing the geometry of a scene as a collection of piecewise smooth and multiply-connected surfaces, the standard in the so-called Marrian approach to Vision [15], is a gross oversimplification that excludes co-dimension one structures (wires), diffuse structures (clouds), and even looking closely at some surfaces (*e.g.*, fabric, hair) it is not clear that there actually is a surface at all, depending on the scale of observation. So, the description of a scene as a configuration of surfaces, although useful in some applications such as robotic navigation, has no more ontological value than the parameters of a trained model, since the latter can be used to infer the former, but not vice-versa. Once restricted to a collection of reflective surfaces, an inferred scene cannot, for instance, be used to model visibility and illumination artifacts such as transparency, translucency, and inter-reflections. In fact, these are considered “outliers” in traditional 3D scene modeling. They are, on the other hand, captured in the representations we have described.
- **Q: You refer to extrapolation as hallucination, but these are two distinct phenomena! Hallucination makes up stuff that is not there in reality.**
- **A:** So does extrapolation. Given a set of data, any form of extrapolation requires some kind of prior, regularizer, or inductive bias, which in a trained model is embodied by the training set, the architecture, the training loss, and the optimization method. These enable one to fill in details not visible in one scene *using information from other scenes*. In standard ill-posed inverse problems, such extrapolation is fostered by generic regularizers (*e.g.*, minimal curvature, maximum sparsity, etc.), but neither the validity of the prior nor the inductive transfer are falsifiable because they are tasked with imputing information from scenes *other than the one in question*, which does not exist in the latter.
- **Q: In what sense is this a Foundation Model? Foundation Models are supposed to be *homogeneous and task-agnostic* representations that can support any downstream tasks. NeRF Diffusions are neither homogeneous nor task-agnostic.**

- **A:** Optimal representations are functions of the data that are minimal sufficient (known also as Sufficient Invariants [1]). This is a well-defined and defining criterion of a representation. Task-agnosticism, on the other hand, is misleading for the only representation that is truly task-agnostic is the data itself, or any lossless representation of it. Even learning criteria branded as task-agnostic, such as contrastive learning and so-called “self-supervised” learning, are very much task-dependent: The task is implicit in the design choice of transformation, data augmentation, or surrogate loss.
- **Q: Foundation Models are Transformer architectures. NeRFs and Diffusion Models are not Foundation Models.**
- **A:** Transformers enjoy some desirable properties that make them suitable for use as Foundations, since they are Turing Complete, a characteristic needed to represent abstract concepts. However, they are not the only ones. So are RNNs and NeRF Diffusions, as we have shown here by proving that NeRF Diffusions can represent the physical scene, which is an abstract concept.
- **Q: You say that a NeRF cannot represent the scene because it is feed-forward, but an MLP enjoys the universal approximation properties of neural networks, so something is wrong.**
- **A:** The Universal Approximation Theorem [6] states that one can approximate a function *in a compact domain*. Scenes do not live in compact domains! If that was the case, we could discretize them and be done with it.
- **Foundation Models are Large Language Models, how does this relate to LLMs?** Transformer architectures pre-trained as masked autoencoders or predictors, and then fine-tuned, possibly using reinforcement learning machinery, on completed sequences have certainly been successful in NLP. While vectorized tokens encode elements in a finite dictionary, there is nothing unique about them representing (sub-)words in a natural language. The same machinery can be used for visual data, consistent with all the derivations in this paper. A further relation is the LLMs, as foundational models for text, are trained to represent “meanings” which are equivalence classes of sentences [37]. A visual foundation model should be trained to represent “scenes,” which as we saw are equivalence classes of images. The major difference is that language admits a natural discretization, being born discrete. Images, on the other hand, cannot be quantized in a way that easily relates to the underlying scene, as the same object can appear smaller than a pixel or larger than the entire image depending on the relative pose to the viewer.
- **The scene is something we evolve to interact with, this paper does not talk about interaction, multimodality, and all of that. Would any of these ideas survive if we expand the scope to a more realistic setting?**
- **A:** In theory, our derivation pertains to any data modality, so it is general. In practice, however, there are modality-specific characteristics that we did not delve into here. For example, remote sensors such as vision behave differently from contact sensors such as touch, and localized modalities differently from diffuse ones such as smell. Most importantly, *active modalities* where the outcome of inference affect the data acquisition process are clearly essential for the development of cognitive abilities (plans do not have a central nervous system), which we do not explore here.
- **Q: you say that the physical scene is an abstract entity, which is contradictory: If it is physical, how can it be abstract?**
- **A:** We have argued that the “true” scene, sometimes referred to as “real” or “physical” scene, cannot be known. So, we define “physical scene” as one that has some uniqueness properties, so that it can be consistent, if not objective, even if observed by multiple agents, each of which processes the data differently. Once so defined, the physical scene is a function of the data stored in the memory of a model, which implements a computable function, that embodies an abstract concept. In general, not just our definition, but any meaningful notion of scene that is inferred from data is an abstract concept,

and attempts to define some sort of objectivity inevitably turn into tautologies, a point eloquently explained by Koenderink in his critique of the objective, or so-called Marrian, account [15] as well as by Russell [10].

- **Q: you also say that physical laws are abstract entities. How can that be?**
- **A:** The laws of physics are a human construct, expressed in human language, that live inside the human brain and are communicated through abstract symbols. These symbols describe relations among quantities, such as $F = ma$, that are not “real” but rather abstract: There are no measurements of force, mass and acceleration that, if plugged into the above equation, would yield the equal sign. That equation, therefore, is not a relation among real measurable quantities, but rather abstractions that can be easily explained and communicated among humans. Trained models may develop their own inner language, as argued in [40], and it is unclear whether that can be “translated” to human natural language so that a representation can “explain” its own version of laws of physics, but all this is beyond our scope here and discussed to some extent in [1].
- **Q: The way in which NeRFs are combined with Diffusion Models is sub-optimal.**
- **A:** There are surely more sophisticated ways of combining NeRFs and Diffusion Models, but this is neither our goal nor the main contribution claimed. Our goal is to test whether even a simple concatenation of the two can suffice to represent a physical scene, in theory - as well as with a modicum of empirical validation. The contribution is not the validation, but the theoretical framing of the question.
- **Q: The definition of scene is trivially satisfied by the images themselves. It is not clear how this representation is meaningful and how it may represent the actual physical scene.**
- **A:** As we point out in Sect. 2.13, indeed a collection of images is a viable representation of a scene, since from them one can infer anything that can be inferred from the scene *given those images*. But this is not a physical scene. As pointed out in the rest of that section, physical scenes are *equivalence classes of scenes*, where the equivalence classes are defined by the *presentations* of the underlying physical scene, if it exists.
- **Q: Why do you need to know the pose? Does the fact that the NeRF is built with posed images invalidate the theory?**
- **A:** In theory, pose is not needed to learn the plenoptic function, as it can be factored out by the model given sufficient training data. In practice, we already know that *a sparse attributed point cloud is minimal sufficient for localization* so long as (a) the attributes (a.k.a. local features descriptors) are sufficient to establish correspondence for a sufficiently large collection of pairs of views, and (b) the number of correspondences is sufficient to define a reference frame (which in turn defines pose) despite violation of the three conditions under which correspondence is possible, which is co-visibility (occlusion), Lambertian reflectance and constant illumination. So, we can simplify the training of the NeRF, by reducing sample complexity, simply bypassing the complex optimization involved in factoring out pose simply by estimating it and inputting it along with the images. If pose is inaccurate, the resulting error will generate a bias, which averages out across samples and is manifest in visible artifacts.
- **Q: What if the scene is not static? Is there an assumption that the scene is immutable across the sensor measurements?**
- **A:** Yes, the scene is defined as what persists among different views. If a scene contains multiple moving “objects” (which we have not defined here as that entails some complexities), each object represents a scene of its own. If objects are simply-connected surfaces supporting Lambertian reflection, this can be done easily. If objects are translucent, transparent, or reflective, there are global dependencies that complicate the analysis, which we defer to future work.