Incorporating Physics into Data-Driven Computer Vision

Achuta Kadambi[1*], Celso de Melo[2], Cho-Jui Hsieh[1], Mani Srivastava[1] and Stefano Soatto[1]

[1*]University of California Los Angeles, 420 Westwood Plaza, Los Angeles, 90095, CA, USA.
[2]DEVCOM U.S. Army Research Laboratory, 12015 Waterfront Drive, Los Angeles, 90094, CA, USA.

*Corresponding author(s). E-mail(s): achuta@ee.ucla.edu;

Abstract

Many computer vision techniques infer properties of our physical world from images. While images are formed through the physics of light and mechanics, computer vision techniques are typically data-driven. This trend is mostly driven by performance: classical techniques from physics-based vision often do not score as high in metrics, compared to modern deep learning. However, recent research, covered in this perspective, has shown that physical models can be included as a constraint into data-driven pipelines. In doing so, one can combine the performance benefits of a data-driven method with advantages offered from a physics-based method, such as intepretability, falsifiability, and generalizability. The aim of this Perspective is to provide an overview into specific approaches of how physical models can be integrated into artificial intelligence (AI) pipelines, referred to as physics-based machine learning. We discuss technical approaches that range from modifications to the dataset, network design, loss functions, optimization, and regularization schemes.

Keywords: Computer Vision, Physics-based Vision, Machine Learning

Modern approaches in computer vision are starting to combine insights from machine learning techniques and physical models. This hybrid approach is referred to as *physics-based learning* (Fig. 1). Computer vision has a special, inherent link to physics, compared to other forms of artificial intelligence (AI) like language that draw primarily from symbolic entities. In particular, many vision techniques infer properties of the physical world from images; and image formation is a process that can be formalized by physical laws. For example, 3D vision involves inference of scene geometry by leveraging physical models that describe how real world points project to virtual camera planes [1, 2]. Video-based computer vision, such as ego-motion control [3–5] leverages the physics of motion to predict states of dynamic agents. The physics of motion takes many forms in computer vision, from a rigid-body described by a trajectory (i.e. group of rigid transformations in 3D space), to complex deformations described by partial differential equations [6–8]. Even semantic tasks like object recognition involve physics. Our semantic notion of an object can be seen as a physical surface surrounded by a medium [9], capable of independent physical motion from a surrounding scene, with geometric (e.g., proximity, shape similarity) [10, 11], photometric (e.g., material similarity), or dynamic (e.g., relative motion) [12] features.

Having described this close link between physics and the foundations of computer vision, one would expect vision algorithms to heavily incorporate physical knowledge. Though physics and vision algorithms are tightly coupled in recent literature, this is a relatively recent development. It is fair to say that physics has not been the focus of the past decade of computer vision. Machine learning has been the focus. Even longstanding problems in vision that have close ties to physical equations are now being addressed with a data-driven approach. Consider the problem of shape reconstruction. The problem was previously addressed with traditional techniques of light transport [18], and now researchers have demonstrated better results when using a neural network [19]. However, while data-driven performance can be superior to a physical model alone, there are problems with a data-driven approach. A neural network is not guaranteed to avoid predictions of shapes or objects that are physically implausible. For example, a neural network for 3D reconstruction will hallucinate detail that is below the resolvable limit of a stereo sensor. Since we know this is not resolvable by a camera, physics would inform us *a priori* that this prediction could be a hallucination. Quantifying the worst case error of a data-driven approach is intrinsically hard due to the inductive hypothesis implicit in data-driven methods. While theoretical machine learning research aims to guarantee neural network performance by bounding error (referred to as generalization bounds [20]), such bounds are only valid under assumptions that cannot be validated in reality, for instance that the finite training data and yet-unseen test data be drawn from the same unknown distribution.

For such reasons, a key question that is being asked is how do we incorporate physics into data-driven pipelines? The motivation is clear: physics and data-driven techniques have complementary strengths and weaknesses, so perhaps the combination will obtain the best of both worlds. Physics can offer interpretable steps and the potential to generalize with limited data, but can be too idealized to describe real-world scenarios. Data-driven methods can

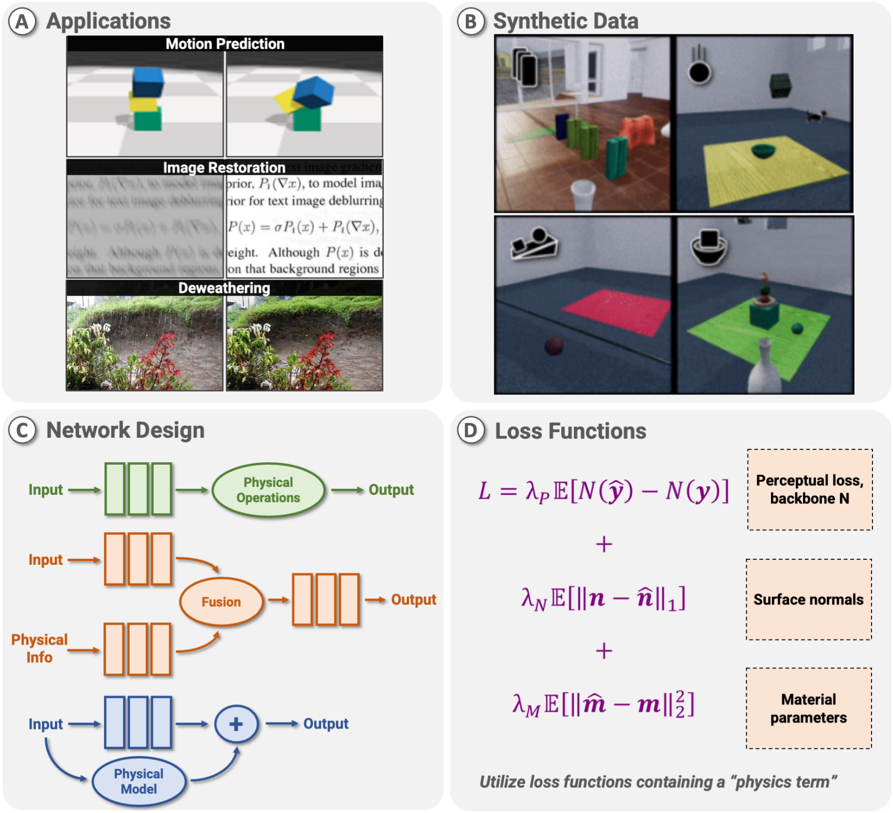
A Applications
Motion Prediction
Image Restoration
prior, $P_l(\nabla x)$, to model ima
rior for text image deblurring
$P(x) = \sigma P_l(x) + P_l(\nabla x)$,
eight. Although $P(x)$ is d
on that background regions
Deweathering
B Synthetic Data
C Network Design
Input
Physical Operations
Output
Input
Fusion
Output
Physical Info
Input
+
Output
Physical Model
D Loss Functions
$L = \lambda_P \mathbb{E}[N(\hat{\boldsymbol{y}}) - N(\boldsymbol{y})]$
Perceptual loss, backbone N
+
$\lambda_N \mathbb{E}[\|\boldsymbol{n} - \hat{\boldsymbol{n}}\|_1]$
Surface normals
+
$\lambda_M \mathbb{E}[\|\hat{\boldsymbol{m}} - \boldsymbol{m}\|_2^2]$
Material parameters
Utilize loss functions containing a "physics term"

Fig. 1 Incorporating physics in neural pipelines in modern computer vision: A) Physics-based learning enables a multitude of applications including motion prediction [13], image restoration [14], and deweathering [15]; B) Deep learning networks can become physics-based if trained on synthetic datasets with strong links to physical rules [16]; C) Neural network architectures can incorporate physics as a constraint to the network topology (figure adapted from [17]); D) Differentiable loss functions that incorporate a physical model can be used to regularize neural networks.

return viable predictions when physical models have model mismatch error, but are not interpretable and require large amounts of data. While combining physics and data might be well motivated, the tactical question of how to combine these entities does not have a single answer. A neural network has many components (weights, losses, inputs, outputs, etc.) and there are multiple ways to incorporate physics into neural networks, with differing tradeoffs.

In this Perspective, we discuss modern methods in vision that have successfully incorporated physics into data-driven pipelines. Many of these methods succeed because they take a holistic approach to methods in visual reasoning. Reasoning in computer vision is usually of an inductive form, and

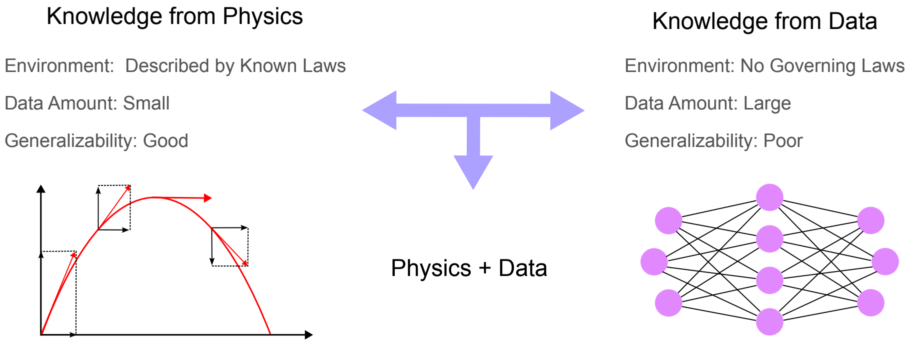

Fig. 2 Illustrating when to approach a problem from a physics-based, data-driven, or hybrid approach. If datasets are small, and environments match physics, then a physics-alone approach makes sense. In contrast, if the dataset size is large and the environments are "real" (deviating from all but the most ideal cases), then a data-driven approach is a better candidate. As we discuss in this piece, many interesting problems benefit from combining both approaches.

these methods incorporate data and physics into the inductive process. Induction is the process of inferring general conclusions from specific information. Any inference process requires biases of some form. Biases can come from design [21] (e.g. choice of an inference or optimization criterion, for instance a segmentation functional or grouping criterion), from physical laws [22, 23] (empirically-validated known constraints), or, as in the modern techniques, from data-driven induction (e.g., the assumption that properties of a finite dataset are shared by the entire distribution of possible data to be measured in the future). Critically, the inductive process does not need to be purely based on physics or data alone. Given where we are as a species, we do not need to learn everything from scratch, so the question arises of how to best make use of verified physical laws in visual inference. As the "why" has been discussed in these introductory paragraphs, the remainder of this piece focuses on "when" and "how" to incorporate physics into data-driven vision pipelines. In particular, Section 1 discusses "when" a problem might merit a physics-based learning approach. Sections 2, 3, 4 focus on the "how" and discuss specific physics-based AI tactics that pertain to datasets, architectures, and loss functions, respectively.

1 When to Use Physics-Based Learning

A first question that this piece addresses is when to incorporate a combined approach of physics and learning (refer Fig. 2). Learning here specifically refers to inductive learning; the process by which a learner or learning algorithm elucidates generalizable rules or functions from a specific set of examples or data. In vision, the data collected by sensors like cameras are inherently lower dimensional than the real world processes they attempt to observe. As such, the data-driven inductive possibilities are assumed to be very large. In contrast, physics-based induction uses a first-order, idealized model that returns a smaller set of inductive possibilities. Therefore, physical laws may be used as an additional inductive bias to reduce the set of generalizable functions provided by a learning algorithm operating on data driven bias only, e.g., by pruning or regularizing any clearly unreasonable solutions. Inductive bias refers to a set of assumptions or rules that the learner uses to predict outputs of given inputs that it has not encountered (i.e. at test time). Such a hybrid approach is known as physics-based machine learning.

Let us return to the question of when to adopt a physics-based learning approach. Consider two extreme cases. In the first case, an inference problem is posed that can just be solved with physics alone, e.g., solving for video tracking of particle motion in an idealized setting. If the accuracy demands of the problem are met with physics alone, the problem should be solved with physics alone. In a second case, a problem can have a negligible relationship to physics, unquantifiable by any form of physical model - such a problem should be solved with data alone. Neither of these two cases are therefore suitable for physics-based learning. However, tasks with partially predictive forward and inverse problems e.g. including but not limited to object recognition in degraded visual conditions [24], super-resolution of satellite imagery [25], system failure prediction [26], are of a third case. These problems lie in a space where physical models are inexact or physical parameters for the models are unknown. In this case, we are better positioned to incorporate this model as an inductive bias, rather than trusting the network to relearn an alternate version of the physical model. A summary of these different paradigms are illustrated in Figure 2).

Therefore, a scenario where physics-based learning should be considered is one where the physics alone is meaningful but, by itself, does not optimally address the inference problem. In particular, there are at least three key considerations one must make in deciding to use physics-based learning: (1) the goodness of data; (2) the goodness of physics; and (3) the ease of integrating data and physics together. The next paragraph outlines technical approaches to assess the "goodness" of data and physics.

There are a few ways to assess the goodness of data with respect to physics. Consider in a first case where the physical model alone can predict the desired task output: then we recommend the use of task performance as an assessment metric. Concretely, the "goodness of data" can be assessed through metrics of task performance using a data-driven approach and compared to the "goodness of physics" by assessing the same metrics of task performance on the physical model alone. While performance metrics are important, one should also consider that the types of errors that a data-driven and physical approach could be different. For example, in deep learning based stereo, one may observe that a physics-based stereo method does not recover fine detail, while a data-driven method is able to superresolve and hallucinate details. Since the data-driven method and physics-based method have different behaviors on task performance, the fusion of a deep learning based stereo with physics-based stereo can be well motivated. However, what about a second case where the physics cannot predict the entire task output: is it still possible to assess the relative quality of physics and data? While end-to-end task output might not be as straightforward to use, one can appeal to *representation probing* where the latent space in a data-driven model is regressed to see if it can predict physics. A third option is to appeal to *intermediate task behavior*, where the performance of a data-driven method is evaluated versus physical models on an intermediate output that physics can predict, which may not be the final task.

Having discussed two conditions (the "goodness" of data, and of physics), we turn to a third condition—the ease of integrating a given physical model with data. A first remark is that integrating physics is easier if the physical model is itself tractable. A tractable model is useful not only for intepretability, but it also enables one to convert machine learning problems from a supervised learning to a self-supervised learning problem, as in the case of deep learning from monocular depth estimation [27–30]. In such examples, a stereo pair is used for data collection, but only one camera is used in the machine learning inference since it is monocular depth estimation. For this case, the problem does not require annotation of data and is self-supervised. Another example of incorporating physics and learning together is when a physical model does not directly predict the inference output, but can prune unreasonable solutions. For example, an object tracking task of dynamic agents like moving vehicles is not described exactly by a physical model: behavioral intent of the driver plays a large role in the possible dynamics. However, even this situation, physics can be used to prune unreasonable solutions. For example, if an object tracker estimates that the vehicle moves from two locations that are further apart than a vehicle's achievable speed would allow, then it can be flagged as a model violation. Yet another type of relation between physics and data pertains to the representation space of an AI pipeline, e.g., in probing a neural representation to see if the physics can be decoded from the latent space. In summary, the section takeaways are: (1) the choice of using a physics-based learning method depends on the quality of physics and data in the problem; and (2) there are specific tactical considerations to assess the value of physics and data in a problem setting.

2 Incorporating Physics into AI Datasets

A first tier of incorporating physics into AI pipelines is to modify the dataset. Even an ordinary neural network can become "physics-based" if the training data used has a strong link to physics. This could be done through synthetic and/or real data. In particular, synthetically generated datasets, where the synthesis is constrained by physical laws (e.g., physics-based engines) helps more efficiently focus the data distribution around the feasible set of data, although it may not cover the tails of the distribution due to oversimplification of synthetic engines. This points to a complementary statement, where coarse behaviors can be captured by synthetic datasets, and fine nuances by raw data.

Consider training an object tracker on two different datasets scenarios. The scenario consists of simulated data of moving pedestrians and cars whose motion is dependent on laws of physics and traffic laws. While this is not a real-world scenario, the concocted, simulated example is dependent on the rules of physics, and neural networks have been shown to implicitly learn approximations of these rules. Now, consider a second scenario of real data of moving pedestrians and cars in a chaotic city. The laws of physics no longer directly predict the motion of pedestrians and cars, as the motion trajectory is not one of a billiard ball, but an autonomous agent that can decide its motion path, based on human behavior and psychology.

However, even in the second scenario, there are some base rules of physics that do carry over (e.g. biophysics dictates that the speed of pedestrians cannot be more than 25 miles per hour). It would be useful to force a network to learn these laws, because many prediction errors we see on real-world object detectors are easily flagged *post-facto*, because of their physical plausibility (i.e. a pedestrian suddenly disappears from the face of the earth, or re-appears further away in a scene than the maximum mobility of a pedestrian would permit).

Flagging prediction errors *post-facto*, is suboptimal: in a deployed model it would be akin to noting the occurrence of an accident after it happens. For this reason it is useful to concoct physics-driven datasets that can be blended with real data to improve AI tasks. For example, [31] used physical models of object collision and intersection to create Stilleben, a framework for generating realistic cluttered scenes for the task of semantic segmentation. Similarly, [32] used UETorch, a version of the popular Unreal game engine with PyTorch incorporated into the game loop, to train a model to predict whether a tower of blocks would fall over and yet other work [33] incorporates a physics engine into a generative model to be able to accurately predict object velocities based on the objects' physical parameters. Other approaches include [34, 35]. These approaches rely on highly effective pipelines for synthetic data augmentation [36].

A future frontier of the field is in increasing the (optical) realism of physically-rendered datasets [37]. The field of physics-based rendering aims to represent the physical properties of light as it travels through a scene. Fortunately, raytracers and other forms of renderers are able to render scenes in accordance with physical laws. Recent approaches known as differentiable rendering, covered in [38], discuss how the forward raytracers are now differentiable, enabling one to optimize scene parameters with respect to visual outputs. This has been extended to more advanced scene physics, for example, beyond photometry research like [39, 40] enable scene understanding in context of polarized light. There are specific approaches that use differentiable rendering like [41] that enable robust estimation of material properties of objects in a scene given a sparse set of views as input. While many of these works developed their own rendering methods, others have used Unity [42, 43], Unreal Engine [44, 45], or other game engines [46], which employ physics-based rendering techniques. Using game-engine rendered synthetic data has allowed many of these works to excel at many vision tasks, such as object detection [44], object tracking [42, 45], or semantic segmentation [43, 46]. [47] developed a physics based model to generate highly realistic faces with blood flow characteristics, which provide robust synthetic data. The use of physical engines can be used beyond the creation of data alone as a way to infer invisible quantities in an image. Zhu et al. [48] demonstrate the inference of forces and pressures—quantities not visible in an image—during human object interactions through physically based simulation. In addition, physically based simulation can be used for other domains, such as learning whether acoustic sensing in a 3D environment can help navigation [49] or learning policies for object tracking with unseen objects, nuisance objects, etc [50].

Despite these advances, there remains a domain gap in how synthetic data maps to real data, underscoring the need for generative models that are even more attuned to real-world physics. Fortunately, there is progress in reducing the domain gap between the simulation and real world, through techniques like domain randomization [51] or a related term, environment augmentation [50, 52]. The basic idea of these techniques is to perturb the generation process of synthetic data, such that the perturbations assist with generalizing to real data. Despite the challenges that need to be overcome to use synthetic data, the use of physically realistic generative models is poised to be an impactful area of research that draws from vision, graphics, and machine learning communities. In summary, the section takeaways are that: (1) datasets can be simulated using known physical lows; (2) AI models trained on this data will be inductively biased toward these laws; and (3) simulation engines exhibit a domain gap (between the real and synthetic worlds) that must be minimized.

3 Incorporating Physics into Network Architectures

A second tier of incorporating physics into AI is to incorporate physics into the inference function. Modern inference functions are deep learning models, and hence this section of the Perspective will focus on incorporating physics into deep learning architectures.

Coupled with recent advances in improving interpretability of deep learning models, various techniques have emerged to incorporate physics and learning together. One technique is known as residual physics, which (as the name suggests) aims to use deep learning to learn the null space of what physics cannot predict. A trivial, data-driven solution is to input video frames into a convolutional neural network to predict trajectories. However, this would be susceptible to the inaccuracies of a data-driven only approach (e.g. requiring large amounts of data, predictions that can grossly violate laws of physics and so on). In the residual physics school of thought (Fig. 3), one may note that

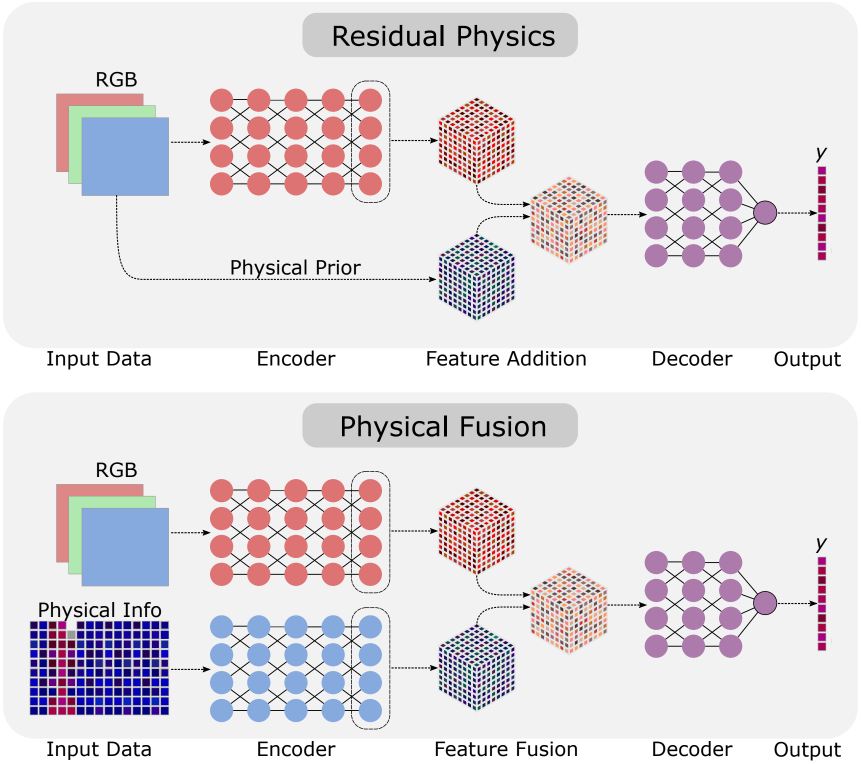

Fig. 3 Two techniques to incorporate physics into machine learning pipelines. (top) Residual physics is an architectural choice where the neural network is geared to predict the residual from the physical model. (bottom) Physical fusion is where physics is treated as a multimodal input to a deep learning model. Late or early fusion can be used to combine features from data and physics.

simple physics, i.e. a parabola equation, can predict the coarse motion arc of the ball. One can then create a skip connection between the parabola prediction and the neural network output. Now the neural network only needs to predict the residual caused by model mismatch in the real data and the simple physical prior of a parabola fit, e.g., air resistance, spin, etc. Many techniques leverage residual physics. For example, [53] use residual physics to teach a robot named ‘TossingBot’ to grasp arbitrary objects from unstructured bins and to throw them into target boxes. The residual learning is employed to predict throw release velocity. TossingBot achieves 85% throwing accuracy. In addition, [54] model uncertainty as residuals for the task of simulating planar pushing and ball bouncing. Further, [55] combine residual physics with neural networks for the task of predicting action effect from sensory data.

Residual physics is not the only way to incorporate physics into deep learning architectures. Indeed, for many problems, residual physics is perhaps not even the best architecture; for example, it requires a fairly accurate physical model to begin with, so that the residual can be bounded to a small norm.

In cases where the physics is a weaker predictor of the output, it might be useful to study a second approach, known as physical fusion, shown in Fig. 3. In this technique, the physical prediction is provided as an input feature (in contrast to residual learning where it was skip connected directly to the output). One can think of the physical prediction as multimodal data, and the network branches into multiple streams that eventually merge to predict the output. This enables physical fusion to be useful in cases where the physics itself is inaccurate and needs to be transformed in a non-linear way before it can be merged into a meaningful representation. As a concrete example, consider the Shape from Polarization [56, 57] problem in computer vision. The goal is to estimate the surface normals of an object given photographs of the scene through different polarizer angles. The relationship between polarization data and shape is a very complex physical model with many unknown constants (like the refractive index, and surface specularity). Therefore, the state of the art methods that use deep learning for shape from polarization incorporate some form of physical fusion by concatenating approximate physical predictions with a dataset [19, 58]. Other work has used physical laws in the form of rule representations as a second encoder branch, where the first encoder branch is a pure data-driven encoder. These are then stochastically concatenated via a control parameter alpha that regulates the strength of the rule on the output. Yet when alpha is fixed prior to training, the trained model cannot operate flexibly based on how much the data satisfies the rule, and therefore rule strength is not adaptable to target data at inference if there is any mismatch with the training setup. Recent work has shown by removing this predetermined constraint on alpha, a higher rule verification ratio, and thus more reliable predictions, can be achieved [59]. Here the rule verification ratio is the fraction of output samples that satisfy the rules. Operating at a better verification ratio could be beneficial, especially if the rules are known to be always valid, as in physics.

While pure deep learning methods are currently used today to attempt answers to scene-related questions such as "where" and "what" an object is, scene understanding of shape, reflectance, and lighting can be improved by incorporating physical priors [60]. The process of achieving these components through intrinsic image decomposition can yield solutions to intricate problems where the "ground truth" is not always available and unsupervised learning with physics-based constraints dominates [61–63]. [64] used the superposition of light to decompose an image with multiple illuminants into separate light-source specific scenes. Learning how light affects an image leads to applications in relighting, where the detected lighting can be replaced with a new source in a different location and color spectrum [65]. Other applications include finding hemoglobin and melanin concentrations on the face through the combination of intrinsic image decomposition and molecule reflectance spectrum modeling [66]. While the reconstruction problem is commonly applied to natural images, the reverse problem of rendering is also inherently physics-based [38, 67–70]. In summary, the section takeaways are that: (1) neural architectures have

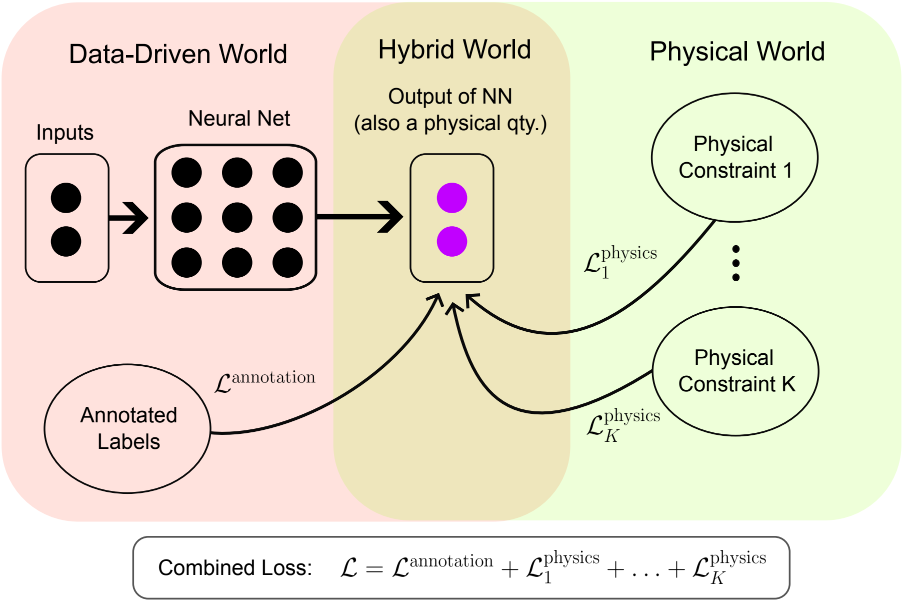
Data-Driven World
Hybrid World
Physical World
Inputs
Neural Net
Output of NN
(also a physical qty.)
Physical Constraint 1
$\mathcal{L}_1^{\text{physics}}$
Physical Constraint K
$\mathcal{L}_K^{\text{physics}}$
Annotated Labels
$\mathcal{L}^{\text{annotation}}$
Combined Loss: $\mathcal{L} = \mathcal{L}^{\text{annotation}} + \mathcal{L}_1^{\text{physics}} + \ldots + \mathcal{L}_K^{\text{physics}}$

Fig. 4 Combined loss functions that use both data-driven annotations and physical constraints. When outputs of a traditional deep learning model are physical quantities, this last, output layer lives in a hybrid world. A compelling case for physics-based learning is made: it is easy to place a loss on the output layer that is based on annotated labels and physical models, as shown in the figure.

emerged that incorporate physics as a constraint or inductive bias; (2) two common example architectures are physical fusion and residual physics; and (3) the choice of architecture is based on factors that include the relevance of the physical model.

4 Incorporating Physics into Network Loss Functions

Related to the previous tier of modifying the neural network topology, a third tier of incorporating physics into deep learning is to incorporate physics into the loss function. When the physical model is known, it can be incorporated into the loss function as a form of regularization. An example is shown in Fig. 4, which involves a data-driven annotation loss and additional loss terms from physical constraints. A few general trends are observed: (1) the loss functions are inspired by well-defined physical priors; (2) these physical priors are often highly domain-specific; and (3) the loss functions are differentiable to enable gradient based learning. If the ground-truth physics is not in a differentiable form, a relaxation to a differentiable function can be used. We will now illustrate a few examples, drawing from diverse tasks in computer vision.

For example, consider the task of vision in bad weather [71]. In this subfield, one goal is to recover a sharp image (e.g. of an outdoor scene) given an input image (which may be corrupted by haze). Since such adverse weather is characterized by the physics of light transport and scattering, we often see differentiable expressions incorporating the physics of light transport and scattering making their way into neural loss functions. For example, [72] proposes a new edge-preserving loss function to enable accurate estimation of the transmission map for dehazing. Loss functions evolve over time, as [73] uses different physics-based priors in the loss formulation to enable synthetic to real transfer of dehazing models. Incorporating physics into the loss function is not limited to weather problems. The task of shadow detection and removal also sees tangible benefits from physics-based loss function design. For example, [74] uses an adversarial shadow attenuation model to improve shadow detection; the shadow-attenuation model relies on a physics-inspired loss incorporating shadow-domain knowledge. Another method in the same area, introduced in [75] uses physics-based chromaticity, boundary smoothness and perceptual features for single image shadow removal. Human body pose estimation is another area in computer vision that leverages physical priors. Various works incorporate the physics of the human body into the supervision for a network, via loss functions on reprojection [76] or joint pose optimization that are combined with data-driven losses [77].

Physics-based loss functions also find significant use in computational imaging tasks as well. For the purpose of positron attenuation correction in computed tomography (CT) imaging, [78] proposes a novel line integral projection loss, consistent with attenuation physics, that leads to improved reconstruction. Other authors have [79] proposed using translation-invariant loss functions for the task of Non Line-of-Sight correlography. And in lensless microscopy, [80] reconstruct phase by fitting the network weights to the captured intensity measurements. Instead of optimizing phase directly, the network optimizes the angular spectrum representation of the measurement in the object plane, allowing an unsupervised training setup. These diverse imaging setups each have their own ad-hoc loss function setups, but the common theme of having a closed-form, differentiable expression that encapsulates domain knowledge is a cross-cutting theme in this area. Looking ahead, much of the future work lies in finding expressions that are both physics-based and yet also differentiable. In cases where this is not always possible, we expect that future work will find relaxations, or use learning to set the parameters of a simpler, differentiable model. In summary, the section takeaways are that: (1) loss functions can incorporate a physical model to regularize a neural network; (2) physics-based loss terms should ideally be differentiable; and (3) if the physics is in a form that does not admit a differentiable loss, then a physically approximate loss that is differentiable can be developed.

5 Future Outlook and Conclusions

The integration of deep learning methods with physics introduces an opportunity to better understand and predict noisy complex natural physical systems.

As discussed here, the integration in these hybrid systems can occur at various levels, from the training data to novel network architectures and loss objectives. As reviewed here, these methods have already shown much promise in enhancing performance in a multitude of forward prediction tasks - object tracking, motion prediction, physical consequences of robot actions, etc. - and inverse problems - scene de-weathering, super-resolution reconstruction of remote imagery, inverse 3D rendering, and more.

An additional direction that is perhaps a few years out lies in unsupervised discovery of physics from visual scenes. We discussed earlier in the piece the work from many authors who have used known physical relationships to recover parameters or directly infer a desired output. However, in some problems one might not have sufficient knowledge of either the underlying physical law or its parameters. This unknown-unknown problem is known as distilling physical laws from data. Physical laws are a human construct, expressed in human language, while recent work with large-scale neural networks hypothesizes the emergence of an "inner language," separate from the human language in which they are trained [81]. A network may then encode physical laws implicitly already, in a language that may not be interpretable by humans. It can be shown that abstract concepts, such as laws of physics, can be finitely represented by a neural network, and are in principle learnable, but external observers cannot know if and when such a concept has been positively encoded [23], although the hypothesis can be falsified. Work in this area is nascent [13, 82–85] and mostly confined to limited settings with relatively simple physical laws for the moment.

The methods described in this Perspective will also play a central role in enabling a next generation of deep neural networks that learn more like biological systems [36, 86–88]. Humans are able to acquire rich internal representations of the physical compositionality of the world by interacting, multimodally and continuously, with objects [89, 90]. By having the ability to reason about the physical properties of the world, as described here, it may become possible to develop novel neural network architectures that are able to interpret scenes by decomposing objects into their physical properties (e.g., shape, surface normals, color) [91], and enabling robust generalization of the learnt knowledge to novel tasks [92].

As this article is being written, modern large language models (LLMs) are exhibiting a remarkable ability to "reason" about many topics, and this includes physics. For example, a recent LLM has shown an ability to outperform the average human test-taker on the Advanced Placement (AP) Physics test, used in the United States [93]. This exciting result should be tempered with the caveat that LLMs cannot learn completely new concepts that are not in their training data [94], and suffer from hallucinations when trying to extrapolate beyond the training data. However, since an LLM is inherently a learning-based method, the ideas in this piece of physics-based learning can be used in a similar fashion as has been discussed to incorporate physics into LLMs. This includes specific ways to incorporate physics into datasets (Section 2), architectures (Section 3) or loss functions (Section 4).

The field of physics-based deep learning provides a path to integrating critical physics knowledge for many visual domains, and also opens the door to novel learning paradigms that will enable a new generation of applications.

Author Contributions:. All authors contributed to the ideas in the manuscript. A.K. took the lead in writing the manuscript. S.S. and C.M. had a supporting role in writing the manuscript. All authors proofread the manuscript.

Competing Interests. A.K. is an employee, receives salary, and owns stock in Intrinsic (an Alphabet Company); and is a co-founder and owns stock in Vayu Robotics. C.M. has no conflicts to declare. C.H., M.S. and S.S. hold employment, draw salary from, and hold stock in Amazon.

Acknowledgments. The authors thank members of the Visual Machines Group for feedback and support, as well as Parth Patwa, Yunhao Ba, Henry Peters, Adnan Armouti, Howard Zhang, Ellin Zhao, Shijie Zhou, Sasha Vilesov, Pradyumna Chari, Zhen Wang, Aditya Gupta, Damon Conover, Akash Singh and Alex Wong for technical discussions, contributions, and pointers to references for this manuscript. This research was partially supported by ARL W911NF-20-2-0158 under the cooperative A2I2 program. A.K. was supported by an NSF CAREER award IIS-2046737, Army Young Investigator Program (YIP) Award, and DARPA Young Faculty Award (YFA).

References

[1] Thapa, S., Li, N., Ye, J.: Dynamic fluid surface reconstruction using deep neural network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21–30 (2020). https://doi.org/10.1109/CVPR42600.2020.00010

[2] Schweri, L., Foucher, S., Tang, J., Azevedo, V.C., Günther, T., Solenthaler, B.: A physics-aware neural network approach for flow data reconstruction from satellite observations. Frontiers in Climate **3**, 656505 (2021)

[3] Zhao, B., Huang, Y., Wei, H., Hu, X.: Ego-motion estimation using recurrent convolutional neural networks through optical flow learning. Electronics **10**(3), 222 (2021)

[4] Zhou, W., Zhang, H., Yan, Z., Wang, W., Lin, L.: Decoupledposenet: Cascade decoupled pose learning for unsupervised camera ego-motion estimation. IEEE Transactions on Multimedia (2022)

[5] Li, W., Zhang, X., Wang, Z., Guo, S., Qiu, N., Li, J.: Dynamic registration: Joint ego motion estimation and 3d moving object detection in dynamic environment. arXiv preprint arXiv:2204.12769 (2022)

[6] Frazzoli, E.: Robust hybrid control for autonomous vehicle motion planning. PhD thesis, Massachusetts Institute of Technology (2001)

[7] Frazzoli, E., Dahleh, M.A., Feron, E.: Real-time motion planning for agile autonomous vehicles. Journal of guidance, control, and dynamics **25**(1), 116–129 (2002)

[8] Goerzen, C., Kong, Z., Mettler, B.: A survey of motion planning algorithms from the perspective of autonomous uav guidance. Journal of Intelligent and Robotic Systems **57**(1), 65–100 (2010)

[9] Gibson, J.J.: The perception of visual surfaces. The American journal of psychology **63**(3), 367–384 (1950)

[10] Latecki, L.J., Lakamper, R.: Shape similarity measure based on correspondence of visual parts. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(10), 1185–1190 (2000)

[11] Mokhtarian, F., Abbasi, S.: Shape similarity retrieval under affine transforms. Pattern Recognition **35**(1), 31–41 (2002)

[12] Raytchev, B., Hasegawa, O., Otsu, N.: User-independent gesture recognition by relative-motion extraction and discriminant analysis. New Generation Computing **18**(2), 117–126 (2000)

[13] Li, Y., Lin, T., Yi, K., Bear, D., Yamins, D.L.K., Wu, J., Tenenbaum, J.B., Torralba, A.: Visual grounding of learned physical models. In: International Conference on Machine Learning (2020)

[14] Pan, J., Dong, J., Liu, Y., Zhang, J., Ren, J., Tang, J., Tai, Y.-W., Yang, M.-H.: Physics-based generative adversarial models for image restoration and beyond. IEEE transactions on pattern analysis and machine intelligence **43**(7), 2449–2462 (2020)

[15] Ba, Y., Zhang, H., Yang, E., Suzuki, A., Pfahnl, A., Chandrappa, C.C., de Melo, C.M., You, S., Soatto, S., Wong, A., *et al.*: Not just streaks: Towards ground truth for single image deraining. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII, pp. 723–740 (2022). Springer

[16] Bear, D.M., Wang, E., Mrowca, D., Binder, F.J., Tung, H.-Y.F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., et al.: Physion: Evaluating physical prediction from vision in humans and machines. arXiv

preprint arXiv:2106.08261 (2021)

[17] Ba, Y., Zhao, G., Kadambi, A.: Blending diverse physical priors with neural networks. arXiv preprint arXiv:1910.00201 (2019)

[18] Atkinson, G.A., Hancock, E.R.: Recovery of surface orientation from diffuse polarization. IEEE transactions on image processing **15**(6), 1653–1664 (2006)

[19] Ba, Y., Gilbert, A., Wang, F., Yang, J., Chen, R., Wang, Y., Yan, L., Shi, B., Kadambi, A.: Deep shape from polarization. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV, pp. 554–571. Springer, Berlin, Heidelberg (2020)

[20] Cao, Y., Gu, Q.: Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 3349–3356 (2020)

[21] Rockwell, C., Johnson, J., Fouhey, D.F.: The 8-point algorithm as an inductive bias for relative pose prediction by vits. In: 3DV (2022)

[22] Lu, Y., Lin, S., Chen, G., Pan, J.: ModLaNets: Learning generalisable dynamics via modularity and physical inductive bias. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 14384–14397. PMLR, ??? (2022). https://proceedings.mlr.press/v162/lu22c.html

[23] Achille, A., Soatto, S.: On the learnability of physical concepts: Can a neural network understand what's real? arXiv (2022). https://doi.org/10.48550/ARXIV.2207.12186

[24] Kilic, V., Hegde, D., Sindagi, V., Cooper, A.B., Foster, M.A., Patel, V.M.: Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection. arXiv preprint arXiv:2107.07004 (2021)

[25] Wang, C., Bentivegna, E., Zhou, W., Klein, L., Elmegreen, B.: Physics-informed neural network super resolution for advection-diffusion models. In: Annual Conference on Neural Information Processing Systems (2020)

[26] Chao, M.A., Kulkarni, C., Goebel, K., Fink, O.: Fusing physics-based and deep learning models for prognostics. Reliability Engineering & System Safety **217**, 107961 (2022)

[27] Zhou, H., Greenwood, D., Taylor, S.: Self-supervised monocular depth

estimation with internal feature fusion. In: BMVC (2021). https://doi.org/10.48550/ARXIV.2110.09482. https://arxiv.org/abs/2110.09482

[28] Klingner, M., Termöhlen, J.-A., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, pp. 582–600. Springer, Berlin, Heidelberg (2020)

[29] Liu, L., Song, X., Wang, M., Liu, Y., Zhang, L.: Self-supervised monocular depth estimation for all day images using domain separation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12737–12746 (2021)

[30] Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2485–2494 (2020)

[31] Schwarz, M., Behnke, S.: Stillleben: Realistic scene synthesis for deep learning in robotics. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 10502–10508 (2020). https://doi.org/10.1109/ICRA40945.2020.9197309

[32] Lerer, A., Gross, S., Fergus, R.: Learning physical intuition of block towers by example. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning, pp. 430–438 (2016)

[33] Wu, J., Yildirim, I., Lim, J.J., Freeman, B., Tenenbaum, J.: Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In: Advances in Neural Information Processing Systems, vol. 28 (2015)

[34] Narang, Y., Sundaralingam, B., Macklin, M., Mousavian, A., Fox, D.: Sim-to-real for robotic tactile sensing via physics-based simulation and learned latent projections. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 6444–6451 (2021). https://doi.org/10.1109/ICRA48506.2021.9561969

[35] Huang, I., Narang, Y., Eppner, C., Sundaralingam, B., Macklin, M., Bajcsy, R., Hermans, T., Fox, D.: Defgraspsim: Physics-based simulation of grasp outcomes for 3d deformable objects. IEEE Robotics and Automation Letters **7**(3), 6274–6281 (2022). https://doi.org/10.1109/LRA.2022.3158725

[36] de Melo, C.M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R.,

Hodgins, J.: Next-generation deep learning based on simulators and synthetic data. Trends in Cognitive Sciences **26**(2), 174–187 (2022). https://doi.org/10.1016/j.tics.2021.11.008

[37] Jalali, B., Zhou, Y., Kadambi, A., Roychowdhury, V.: Physics-ai symbiosis. Machine Learning: Science and Technology **3**(4), 041001 (2022)

[38] Zhao, S., Jakob, W., Li, T.-M.: Physics-based differentiable rendering: From theory to implementation. In: ACM SIGGRAPH 2020 Courses. SIGGRAPH '20. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3388769.3407454. https://doi.org/10.1145/3388769.3407454

[39] Baek, S.-H., Zeltner, T., Ku, H.J., Hwang, I., Tong, X., Jakob, W., Kim, M.H.: Image-based acquisition and modeling of polarimetric reflectance. ACM Trans. Graph. **39**(4) (2020). https://doi.org/10.1145/3386569.3392387

[40] Kondo, Y., Ono, T., Sun, L., Hirasawa, Y., Murayama, J.: Accurate polarimetric brdf for real polarization scene rendering. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020, pp. 220–236. Springer, Cham (2020)

[41] Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

[42] Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtualworlds as proxy for multi-object tracking analysis. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4340–4349 (2016). https://doi.org/10.1109/CVPR.2016.470

[43] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243 (2016). https://doi.org/10.1109/CVPR.2016.352

[44] Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O., Birchfield, S.: Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 7249–7255 (2019). https://doi.org/10.1109/ICRA.2019.8794443

[45] Müller, M., Casser, V., Lahoud, J., Smith, N., Ghanem, B.: Sim4cv: A photo-realistic simulator for computer vision applications. International

Journal of Computer Vision **126**(9), 902–919 (2018). https://doi.org/10.1007/s11263-018-1073-7

[46] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016, pp. 102–118. Springer, Cham (2016)

[47] Wang, Z., Ba, Y., Chari, P., Bozkurt, O.D., Brown, G., Patwa, P., Vaddi, N., Jalilian, L., Kadambi, A.: Synthetic generation of face videos with plethysmograph physiology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20587–20596 (2022)

[48] Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., Zhu, S.-C.: Inferring forces and learning human utilities from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3823–3833 (2016)

[49] Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pp. 17–36 (2020). Springer

[50] Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., Wang, Y.: End-to-end active object tracking and its real-world deployment via reinforcement learning. IEEE transactions on pattern analysis and machine intelligence **42**(6), 1317–1332 (2019)

[51] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 23–30 (2017). IEEE

[52] Sadeghi, F., Levine, S.: CAD2RL: real single-image flight without a single real image. In: Amato, N.M., Srinivasa, S.S., Ayanian, N., Kuindersma, S. (eds.) Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017 (2017). https://doi.org/10.15607/RSS.2017.XIII.034. http://www.roboticsproceedings.org/rss13/p34.html

[53] Zeng, A., Song, S., Lee, J., Rodriguez, A., Funkhouser, T.: Tossingbot: Learning to throw arbitrary objects with residual physics. Trans. Rob. **36**(4), 1307–1319 (2020). https://doi.org/10.1109/TRO.2020.2988642

[54] Ajay, A., Wu, J., Fazeli, N., Bauza, M., Kaelbling, L.P., Tenenbaum, J.B.,

Rodriguez, A.: Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing. In: IROS 2018 (2018)

[55] Kloss, A., Schaal, S., Bohg, J.: Combining learned and analytical models for predicting action effects from sensory data. The International Journal of Robotics Research, 0278364920954896 (2020)

[56] Kadambi, A., Taamazyan, V., Shi, B., Raskar, R.: Polarized 3d: High-quality depth sensing with polarization cues. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3370–3378 (2015)

[57] Kalra, A., Taamazyan, V., Rao, S.K., Venkataraman, K., Raskar, R., Kadambi, A.: Deep polarization cues for transparent object segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8599–8608 (2020). https://doi.org/10.1109/CVPR42600.2020.00863

[58] Zou, S., Zuo, X., Qian, Y., Wang, S., Xu, C., Gong, M., Cheng, L.: 3d human shape reconstruction from a polarization image. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV, pp. 351–368. Springer, Berlin, Heidelberg (2020)

[59] Seo, S., Arik, S., Yoon, J., Zhang, X., Sohn, K., Pfister, T.: Controlling neural networks with rule representations. In: Advances in Neural Information Processing Systems (2021)

[60] Klinghoffer, T., Somasundaram, S., Tiwary, K., Raskar, R.: Physics vs. Learned Priors: Rethinking Camera and Algorithm Design for Task-Specific Imaging. arXiv (2022). https://doi.org/10.48550/ARXIV.2204.09871. https://arxiv.org/abs/2204.09871

[61] Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J.B.: Self-supervised intrinsic image decomposition. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 5938–5948. Curran Associates Inc., Red Hook, NY, USA (2017)

[62] Vamaraju, J., Sen, M.K.: Unsupervised physics-based neural networks for seismic migration. Interpretation **7**(3), 189–200 (2019)

[63] Rupe, A., Kumar, N., Epifanov, V., Kashinath, K., Pavlyk, O., Schlimbach, F., Patwary, M., Maidanov, S., Lee, V., Prabhat, M., *et al.*: Disco: Physics-based unsupervised discovery of coherent structures in spatiotemporal systems. In: 2019 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC), pp. 75–87 (2019).

IEEE

[64] Hui, Z., Chakrabarti, A., Sunkavalli, K., Sankaranarayanan, A.C.: Learning to separate multiple illuminants in a single image. In: Computer Vision and Pattern Recognition (CVPR 2019) (2019)

[65] Nestmeyer, T., Lalonde, J., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5123–5132. IEEE Computer Society, Los Alamitos, CA, USA (2020). https://doi.org/10.1109/CVPR42600.2020.00517. https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00517

[66] Alotaibi, S., Smith, W.A.P.: Biofacenet: Deep biophysical face image interpretation. In: Proc. of the British Machine Vision Conference (BMVC) (2019)

[67] Cai, G., Yan, K., Dong, Z., Gkioulekas, I., Zhao, S.: Physics-based inverse rendering using combined implicit and explicit geometries. arXiv preprint arXiv:2205.01242 (2022)

[68] Halder, S.S., Lalonde, J.-F., Charette, R.d.: Physics-based rendering for improving robustness to rain. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10203–10212 (2019)

[69] Agarwal, A., Man, T., Yuan, W.: Simulation of vision-based tactile sensors using physics based rendering. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 1–7 (2021). IEEE

[70] Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., *et al.*: Advances in neural rendering. In: Computer Graphics Forum, vol. 41, pp. 703–735 (2022). Wiley Online Library

[71] Nayar, S.K., Narasimhan, S.G.: Vision in bad weather. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 820–8272 (1999). https://doi.org/10.1109/ICCV.1999.790306

[72] Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3194–3203 (2018). https://doi.org/10.1109/CVPR.2018.00337

[73] Chen, Z., Wang, Y., Yang, Y., Liu, D.: Psd: Principled synthetic-to-real dehazing guided by physical priors. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7176–7185 (2021). https://doi.org/10.1109/CVPR46437.2021.00710

[74] Le, H., Vicente, T.F.Y., Nguyen, V., Hoai, M., Samaras, D.: A+d net: Training a shadow detector with adversarial shadow attenuation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

[75] Jin, Y., Sharma, A., Tan, R.T.: Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5007–5016 (2021). https://doi.org/10.1109/ICCV48922.2021.00498

[76] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, pp. 561–578 (2016). Springer

[77] Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: Physcap: Physically plausible monocular 3d motion capture in real time. ACM Transactions on Graphics (ToG) **39**(6), 1–16 (2020)

[78] Shi, L., Onofrey, J.A., Revilla, E.M., Toyonaga, T., Menard, D., Ankrah, J., Carson, R.E., Liu, C., Lu, Y.: A novel loss function incorporating imaging acquisition physics for pet attenuation map generation using deep learning. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, pp. 723–731. Springer, Cham (2019)

[79] Metzler, C.A., Heide, F., Rangarajan, P., Balaji, M.M., Viswanath, A., Veeraraghavan, A., Baraniuk, R.G.: Deep-inverse correlography: towards real-time high-resolution non-line-of-sight imaging: erratum. Optica **7**(3), 249–251 (2020). https://doi.org/10.1364/OPTICA.391291

[80] Zhang, F., Liu, X., Guo, C., Lin, S., Jiang, J., Ji, X.: Physics-based iterative projection complex neural network for phase retrieval in lensless microscopy imaging. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10518–10526 (2021). https://doi.org/10.1109/CVPR46437.2021.01038

[81] Trager, M., Perera, P., Zancato, L., Achille, A., Bhatia, P., Xiang, B., Soatto, S.: Linear spaces of meanings: the compositional language of vision-language models. arXiv preprint arXiv:2302.14383 (2023)

[82] Fragkiadaki, K., Agrawal, P., Levine, S., Malik, J.: Learning visual predictive models of physics for playing billiards. arXiv preprint arXiv:1511.07404 (2015)

[83] Chari, P., Talegaonkar, C., Ba, Y., Kadambi, A.: Visual physics: Discovering physical laws from videos. arXiv preprint arXiv:1911.11893 (2019)

[84] Li, Y., Torralba, A., Anandkumar, A., Fox, D., Garg, A.: Causal discovery in physical systems from videos. Advances in Neural Information Processing Systems **33**, 9180–9192 (2020)

[85] Chen, B., Huang, K., Raghupathi, S., Chandratreya, I., Du, Q., Lipson, H.: Automated discovery of fundamental variables hidden in experimental data. Nature Computational Science **2**(7), 433–442 (2022)

[86] Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-Inspired artificial intelligence. Neuron **95**(2), 245–258 (2017)

[87] Marblestone, A.H., Wayne, G., Kording, K.P.: Toward an integration of deep learning and neuroscience. Frontiers in computational neuroscience, 94 (2016)

[88] Bengio, Y., Lee, D.-H., Bornschein, J., Mesnard, T., Lin, Z.: Towards biologically plausible deep learning. arXiv preprint arXiv:1502.04156 (2015)

[89] Battaglia, P.W., Hamrick, J.B., Tenenbaum, J.B.: Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences **110**(45), 18327–18332 (2013) https://www.pnas.org/doi/pdf/10.1073/pnas.1306572110. https://doi.org/10.1073/pnas.1306572110

[90] Spelke, E.S., Kinzler, K.D.: Core knowledge. Developmental Science **10**(1), 89–96 (2007)

[91] Wu, J., Lim, J.J., Zhang, H., Tenenbaum, J.B.: Physics 101: Learning physical object properties from unlabeled videos. In: BMVC (2016)

[92] Bear, D.M., Fan, C., Mrowca, D., Li, Y., Alter, S., Nayebi, A., Schwartz, J., Fei-Fei, L., Wu, J., Tenenbaum, J.B., Yamins, D.L.K.: Learning physical graph representations from visual scenes. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (2020)

[93] OpenAI: GPT-4 technical report. Technical report, OpenAI (2023). https://cdn.openai.com/papers/gpt-4.pdf

[94] Chrupala, G., Alishahi, A., Berg-Kirkpatrick, T.: The science of language modeling. Annual Review of Linguistics **7**, 149–176 (2021)