# CAUSAL FUTURE PREDICTION IN A MINKOWSKI SPACE-TIME

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Estimating future events is a difficult task. Unlike humans, machine learning approaches are not regularized by a natural understanding of physics. In the wild, a plausible succession of events is governed by the rules of causality, which cannot easily be derived from a finite training set. In this paper we propose a novel theoretical framework to perform causal future prediction by embedding spatio-temporal information on a Minkowski space-time. We utilize the concept of a light cone from special relativity to restrict and traverse the latent space of an arbitrary model. We demonstrate successful applications in causal image synthesis and future video frame prediction on a dataset of images. Our framework is architecture- and task-independent and comes with strong theoretical guarantees of causal capabilities.

## 1 INTRODUCTION

In many everyday scenarios we make causal predictions to assess how situations might evolve based on our observations and experiences. Machine learning has not been developed to this level yet, though, automated, causally plausible predictions are highly desired for critical applications like medical treatment planning, autonomous vehicles and security. Recent works have contributed machine learning algorithms for the prediction of the future in sequences and for causal inference Kurutach et al. (2018). One major assumption that many approaches implicitly adopt, is that the space of the model representation is a flat Euclidean space of N dimensions. However, as shown by Arvanitidis et al. Arvanitidis et al. (2018), the Euclidean assumption leads to false conclusions as a model's latent space can be better characterized as a high dimensional curved Riemannian manifold rather than an Euclidean space. Furthermore, the Alexandrov-Zeeman theorem Zeeman (1964); Kosheleva & Kreinovich (2014) suggests that causality requires a Lorentzian group space and advocates the unsuitability of Euclidean spaces for causal analysis.

In this paper, we present a novel framework that changes the way we treat hard computer vision problems like the continuation of frame sequences. We embed information on a spatio-temporal, high dimensional pseudo-Riemannian manifold - the Minkowski space-time - and utilize the special relativity concept of light cones to perform causal inference. We focus on temporal sequences and image synthesis to exhibit the full capabilities of our framework. In summary our contributions are:

- We extend representation learning to spatio-temporal Riemannian manifolds that follow the ideas of the Minkowski space-time while being agnostic towards the used embedding architecture and the prescribed task.
- We introduce a novel utilization of the concept of light cones and use them for convincing frame synthesis and plausible prediction of future frames in video sequences.
- We provide theoretical guarantees about the causal properties of our model and demonstrate a causal inference framework.

## 2 RELATED WORKS

High dimensional Riemannian manifolds for machine learning are utilized by a few major works. Arvanitidis *et al.* Arvanitidis et al. (2018) show evidence that more general Riemannian manifolds characterize learned latent spaces better than an Euclidean space. Their work however, utilizes generators that have been trained under an Euclidean assumption. Contrary to that, Nickel *et al.* Nickel & Kiela (2017) introduce the use of a Poincaré ball for hierarchical representation learning on word embeddings, showing superior performance in representation capacity and generalization ability while employing a Riemannian optimization process. In Nickel & Kiela (2018), Nickel *et al.* extend the previous work to a Lorentzian manifold as this offers improvements in efficiency and stability of the distance function. In this paper we accept the argument made by Nickel *et al.* but extend it as we argue in Section 3 that causal inference requires a Lorentzian group space Zeeman (1964).

Ganea *et al.* Ganea et al. (2018) embed word information on a Poincaré ball and form entailment cones. The authors propose to work with Directed Acyclical Graphs (DAG) and strive for non overlapping cones in a Poincaré ball. In contrast to this, we encourage overlapping light cones in a Lorentzian manifold to model future events.

Sun *et al.* Sun et al. (2015) use a space-time idea similar to ours but interpret the time axis as a ranking rather than as temporal information. Their method is intended for dimensionality reduction and does not generate further samples, or considers causal relationships between sampling points. Finally, Mathieu *et al.* Mathieu et al. (2019) train a Variational Autoencoder (VAE) constrained to a Poincaré ball while also employing the appropriate Riemannian equivalent to a normal distribution as well as Riemannian optimization. We consider this work as the closest related since it is the only approach that has shown good performance in the image domain.

In the Computer Vision focused field of future frame prediction for video sequences, Kurutach et al. (2018) propose the causal InfoGAN which, however, lacks theoretical guarantees of causal abilities. Jayaraman et al. (2019) aims at predicting the probabilistic bottlenecks where the possible futures are constrained instead of generating a single future. Similarly, we are not attempting to predict a single future, rather we predict all plausible futures in a way that naturally enables us to identify all probabilistic bottlenecks; see Section 3. In other works concerned with video continuation, Mathieu et al. (2016); Vondrick et al. (2016a) use CNNs to regress future frames directly, while Villegas et al. (2017a) introduce an LSTM utilizing the difference $\Delta$ between frames to predict motion. Further works include the use of optical flow Liu et al. (2018) or human pose priors Villegas et al. (2017b). The autoregressive nature of these methods results in accumulated prediction errors that are propagated through the frames the further a sequence is extended. Beyond a few frames, these approaches quickly lose frame-to-frame consistency. In order to mitigate these limitations, works like Vondrick et al. (2016b) propose generative models to predict future frames and Tulyakov et al. (2018) offers a generative model that disentangles motion and content. Neither can infer the causal implications of their starting positions.

## 3 THEORETICAL FORMULATION

**Causal Inference** Causal inference refers to the investigation of causal relations between data. There is a rich literature on machine learning and causal inference ranging from association of events to counterfactuals Peters et al. (2019); Pearl (2019). Briefly we observe two equivalent approaches towards causality in machine learning: Structural Causal Models Pearl (2019); Pearl et al. (2016) which rely on Directed Acyclical Graphs (DAG) and Rubin Causal Models Rubin (2005) which rely upon the potential outcomes framework. In this paper we will be focusing on the latter. In the potential outcomes framework as established by Rubin (2005) multiple outcomes $\mathcal{Y}$ of $\mathcal{X}$ are contrasted in order to deduce causal relations between $\mathcal{Y}$ and $\mathcal{X}$. As we will show, our proposed method provides the theoretically guaranteed infrastructure to create a Rubin Causal

Model. In addition, as our method is able to operate in a future as well as a past regime it enables the formation of counterfactual questions, *i.e.*, what would $\mathcal{Y}$ be if $\mathcal{X}'$ had happened instead.

**On the choice of space:** In his seminal 1964 work, E.C. Zeeman Zeeman (1964) makes the case that the causality group $\mathcal{R}M$ that arises from the concept of partial ordering in a Minkowski space-time implies an inhomogenous Lorentz group as the symmetry group of $\mathcal{R}M$. We highlight the explicit mention of Zeeman on the unsuitability of an Euclidean topology to describe $\mathcal{R}M$ due to its local homogeneity, which does not arise in $\mathcal{R}M$. In Kosheleva & Kreinovich (2014) the authors prove that from observable causality we can reconstruct the Minkowski space-time. Hence, we are in a position to argue that the use of a Minkowski space-time for embeddings, which belongs to the inhomogenous Lorentz group, would reinforce causal inference capabilities.

We define our Minkowski space-time to be characterized by the metric of Eq. 1 with the element $-1$ denoting the temporal dimension and $+1$ elements the spatial dimensions. We extend Nickel & Kiela (2018) and argue that the use of the Lorentzian manifold, which coincides with the Minkowski space-time, is both more efficient as an embedding as well as enabling causal arguments,

$$\eta_{\mu\nu} = \text{diag}(-1, +1, +1, +1). \tag{1}$$

**Minkowski Space-Time and Causality:** Mathematically a space can be described by its metric, which defines the way the inner product of two vectors in this space is determined, *i.e.* the way we calculate distances. Consequently, the inner product $\langle ., . \rangle_\eta$ of two vectors $a$ and $b$ in $1 + 3D$ Minkowski space-time can be defined as

$$\langle a, b \rangle_\eta = \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} a_\mu \eta_{\mu\nu} b_\nu = -a_0 b_0 + a_1 b_1 + a_2 b_2 + a_3 b_3, \tag{2}$$

where the coordinate 0 is understood to be the time coordinate.

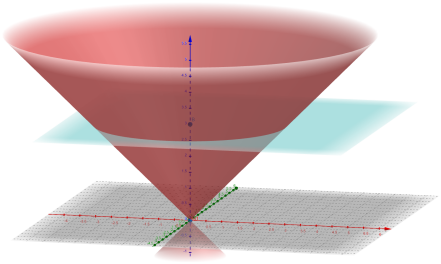
One of the consequences of endowing the latent space with a Minkowski-like metric is the emergence of causality in the system. This property can be more readily seen by employing the concept of *proper time*. Given a manifold $\mathcal{M}$ endowed with a Minkowski metric $\eta_{\mu\nu}$, we define the proper time $\tau$. This is the time measured by an observer following along a continuous and differentiable path $\mathcal{C}(s)$ parametrized by $s \in [0, 1]$ between two events $\{x, y\} \in \mathcal{M}$ such that $\mathcal{C}(0) = x$, $\mathcal{C}(1) = y$,

$$\tau_{\mathcal{C}} = \int_{\mathcal{C}} \sqrt{-\sum_{\mu,\nu} dx_\mu dx_\nu}. \tag{3}$$

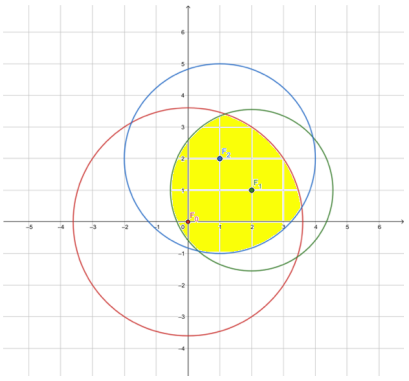
In order to ensure $\tau \in \mathbb{R}$, we require $\sum_i dx_i^2 \leq dx_0^2$, where $i \in 1, 2, ..., d$. Therefore, the rate of change $|\mathbf{dx}|/d\tau$ in the spatial coordinates is capped by the time evolution of the system. In other words, there exists a maximum speed limit which $\mathcal{C}$ must obey at every point. Further, it means that there exist pairs of space-time points $x, y$ which cannot be possibly connected by a valid path $\mathcal{C}$ , lest $\tau \notin \mathbb{R}$. In order to describe this phenomenon we borrow the concept of a *light cone* from special relativity. The set of solution paths $\{\mathcal{C}_0(s)\}$ such that $\mathcal{C}_0(0) = (t_0, \mathbf{x_0})$ and $\tau_{\mathcal{C}_0} = 0$ describe the fastest any particle or piece of information can travel within the system starting from $(t_0, \mathbf{x_0})$. This boundary is known as the light cone, and is such that $\partial\mathcal{R} = \{\mathcal{C}_0(s)\}$, where $\mathcal{R}$ is the causal region of the point $(t_0, \mathbf{x_0})$. Every space-time point $x \in \mathcal{R}$ is said to be within the light cone. As shown by (3), no valid path $\mathcal{C}(s)$ can cross $\partial\mathcal{R}$. Thus, two space-time points can only influence each other if they lie within each other's light cone, that is, if they can be connected by a valid path $\mathcal{C}$. The region $\mathcal{R}$ splits into two disjoint sets: $\mathcal{R}^+$ and $\mathcal{R}^-$. $\mathcal{R}^+$ lies within the future light cone of a particle at time $t_0$, and thus includes all of the points $(t_1, \mathbf{x_1}) \in \mathcal{R}$ such that $t_1 > t_0$. Conversely, $\mathcal{R}^-$ includes the points $(t_2, \mathbf{x_2}) \in \mathcal{R}$ such that $t_2 < t_0$ and characterizes the past light cone of a particle at time $t_0$.

If we have two space-time vectors $x = (t_0, \mathbf{x_0})$ and $y = (t_1, \mathbf{x_1})$ we can describe their relation as *timelike* when $\langle x, y \rangle < 0$, *spacelike* when $\langle x, y \rangle > 0$ and *lightlike* when $\langle x, y \rangle = 0$. A timelike position vector lies within the light cone of a particle at the origin of the system. A spacelike vector lies outside of it, and a lightlike vector lies exactly at its edge. One can then generalize this idea beyond the origin, and thus compute the inner product of the difference between two space-time vectors $x - y \equiv (\Delta t, \Delta \mathbf{r})$ , *i.e.*, $\langle y - x, y - x \rangle = -\Delta t^2 + |\Delta \mathbf{r}|^2$. Hence, when the separation of the vectors $x$ and $y$ is timelike, they lie within each other's causal region. In that case we can argue that there is a path for particle $x$, that belongs in the model that defines the latent world of represented data, to evolve into particle $y$ within a time period $\Delta t$. Thus, by constructing the light cone of an initial point $x$ we can constrain the space where the causally resulting points may lie. We can then see that this mathematical construction of the latent space naturally enforces that the velocity of information propagation in the system be finite, and that a particle can only be influenced by events within its past light cone, *i.e.* the model is causal. By mapping this into a machine learning perspective we argue that in a latent space that is built to follow the Minkowski space-time metric an encoded point can then be used to create a light cone that constrains where all the causally entailed points may be encoded to or sampled from.

**On Intersecting Cones:** A light cone can be constructed with each point of the latent space as its origin. Consider point $x_0$ to be an initial point derived from, for example, an encoded frame $f_0$ from a video sequence: by constructing the light cone $C_0$ around $x_0$ we are able to deduce where the various causally related $x_{0+t}$ points might lie. By setting $t$ to be a specific time instance, we are able to further constrain the sub-space to points that lie inside of the conic section. They are causally plausible results of point $x_0$ within the time $t$. Geometrically, we can visualize this as a plane cutting a cone at a set time. We visualize this in Figure 1a.



(a) Visualization of the emerging structure of a light cone. The intersecting plane at point $z = 3$ signifies the 2-dimensional feature space at time 3. The interior of the cone subspace contains all possible frames given a original video frame at point $z = 0$.



(b) Visualization of the intersecting cones algorithm. The subspace marked in yellow contains the points that are causally related to points $F_{0,1,2}$.

Figure 1: Visual aids of proposed algorithm. Note that for visualization purposes we are exhibiting a $1 + 2$ dimensional Euclidean space rather than a high dimensional Riemannian manifold.

A second point $x_1$ that lies inside the light cone of $x_0$ can be derived from an encoded frame $f_1$. Similar to $x_0$ we construct the light cone $C_1$ whose origin is $x_1$. We then define the conic intersection $CS = C_0 \cap C_1$. Following the causality argument, we deduce that the enclosed points in $CS$ are causally related to both $x_0$

and $x_1$ as they lie in the light cones of both. In addition, by constraining the intersecting time plane, we constrain the horizon of future prediction.

Consequently, we propose Algorithm 1 as a method of future frame prediction using light cones on a Minkowski space-time latent space. We graphically represent Algorithm 1 in Figure 1b.

**Algorithm 1:** Future Prediction using Intersecting Light Cones

**Input :** Frame Sequence $F$ ; Queried Time $T$
**Output:** Predicted Frame
**for** $t < T$ **do**
  $Mf_t \leftarrow MinkowskiEmbedding(f_t)$
  $C_{Mf_t} \leftarrow LightCone(Mf_t)$
  **if** $t > len(F)$ **then**
    $Samples_{Mf_t} \leftarrow sample(C_{Mf_t})$
    $Mf_{t+k} \leftarrow choose(Samples_{Mf_t})$
  **end**
**end**
$CS \leftarrow intersection(C_{MF})$
$f_{out} \leftarrow choose(sample(CS))$
Predicted Frame $\leftarrow Decoder(f_{out})$

**On the Entropy and the Aperture of Cones :** When considering the intersection of the cones in Algorithm 1 it is vital to examine the aperture of the cone at time $T$. For simplicity, we assume that the gradient of the side of the cone is $45^\circ$ for all cones. However, such an assumption implies that each frame and hence each cone evolves with the same speed and can reach the same number of states at a given time. For real world scenarios this is not necessarily true as, for example, the possible states in $t + 1$ for a ball rolling constraint by rails are less than a ball rolling on a randomly moving surface. Hence, the actual gradient of the cone depends on the number of states that are reachable from the state depicted in frame $t$. This quantity is also known as the thermodynamic entropy of the system. It is defined as the sum of the states the system can evolve to. Calculating the thermodynamic entropy of a macro-world system as in a real world dataset is not trivial and we are not aware of any appropriate method to compute this at the time of writing. Hence, we are forced to make the aforementioned assumption of $45^\circ$.

However, given a frame sequence $F$, a set of counter example frames $CF$ and following Algorithm 1 but omitting the sampling steps, it is possible to build more accurate light cones in a contrastive manner. Hence, it is possible to acquire a proxy for the thermodynamic entropy of the system. We note that the proxy can only be accurate to a certain degree as any frame sequence is not able to contain enough information to characterize the full state of the world.

## 4 EXPERIMENTATION

**Training:** Our proposed algorithm is invariant to the method used to train the embedding. In an ideal scenario, we require an encoder-decoder pair that is able to map any image to a latent space and to reconstruct any latent code. For the purposes of this paper's evaluation we have chosen the method by Mathieu *et al.* Mathieu et al. (2019) as our baseline embedding, as it is the only approach that has shown good image domain performance.

Mathieu *et al.* Mathieu et al. (2019) construct a Variational Auto Encoder (VAE) that enforces the latent space to be a Poincaré Ball. We analyze the properties of the Poincaré ball in the supplementary material. It can be shown Nickel & Kiela (2018) that a $n-$dimensional Poincaré ball embedding can be mapped into a

subspace of the Minkowski space-time by an orthochronous diffeomorphism $m : P^n \rightarrow M^n$,

$$m(x_1, ...x_n) = \frac{(1 + ||x||^2, 2x_1, ..., 2x_n)}{1 - ||x||^2},$$

and back with the inverse $m^{-1} : M^n \rightarrow P^n$

$$m^{-1}(x_1, ...x_n) = \frac{(x_1, ..., x_n)}{1 + x_0},$$

where $x_i$ is the i-th component of the embedding vector.

We extend Mathieu et al. (2019) to enforce the embedding to a subspace of the Minkowski space-time by utilizing Eq. 4 and 4. We treat the space's dimensionality as hyper-parameter and tune it experimentally. We establish that the optimal embedding of our data can be achieved in an $1 + 8$ dimensional space *i.e.* 1 time and 8 space dimensions. The model consists of a MLP with a single hidden layer and was trained with the Riemannian equivalent of the Adam optimizer Sun et al. (2015) with a learning rate of $5e - 4$. Training the model with Moving MNIST requires on a Titan RTX Nvidia GPU less than 1 hour.

**Inference:** Our proposed Algorithm 1 is executed during inference as it does not require any learned parameters. We sample from a Gaussian distribution wrapped to be consistent with our Minkowski space-time in a manner similar to Mathieu et al. (2019), details can be found in the supplement. Inference can be performed in about $0.5$ s per intersecting cone in an exponential manner.

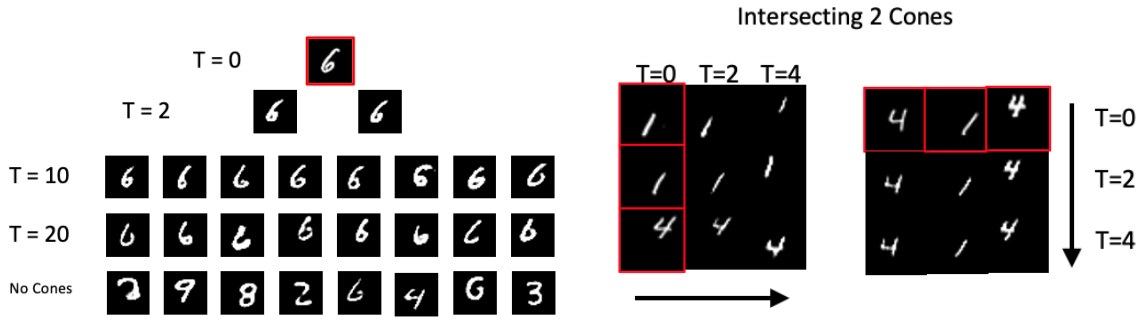
**Dataset:** As a proof of concept we use a custom version of the Moving MNIST datasetSrivastava et al. (2015). Specifically we employ 10.000 sequences consisting of 30 frames each, making a total of 300.000 frames. Each sequence contains a single digit. The first frame is derived from the training set of the original MNIST dataset, while the subsequent frames are random continuous translations of the digit. Construction of the test set followed the same procedure with the first frame derived from the test set of the original MNIST dataset. We created 10.000 testing sequences of 30 frames each. Each frame is $32 \times 32$ while the containing digits range from $18px - 25px$

We further use the KTH action recognition dataset Schuldt et al. (2004) to highlight the real world capabilities of our method. We focus on the walking and handwaving actions and use all 4 distinct directions. Different person identities are used in the train-test split.

## 5 RESULTS

**Experiment 1: Single Cone Image Synthesis** In the first experiment we evaluate the ability of the light cone to constrain the latent space such that samples lying inside the cone are reasonably similar to the original frame. We train our model with 1+8 latent dimensions. Following standard VAE sampling we produce 100.000 random samples using a wrapped normal distribution. As expected, the tighter the imposed time bound is, the fewer samples are accepted. We note that for $t = 2$ only 2 samples were accepted, for $t = 10$ our method accepts $N = 31\%$ of the samples and for $t = 20$, $N = 71\%$. In Figure 2a we exhibit qualitative results for Experiment 1. We note that as the time limit increases we observe higher variability, both in terms of morphology and location of the digits, while the identity of the digit remains the same. This is in accordance with the theory that the "system" would have enough time to evolve into new states. More examples are included in the supplementary material.

**Experiment 2: Intersecting Cones** In the second experiment we evaluate the ability of our algorithm to predict frames by intersecting light cones. There is no unique path a system might evolve in time. Our algorithm does not aim at producing a single future, rather it is able to produce multiple plausible futures.




(a) Samples from Experiment 1.

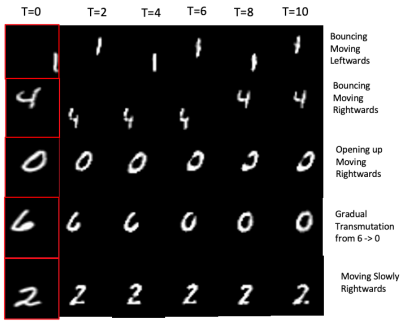
(b) Samples from Experiment 2.

Figure 2: (a): Random sampling was constrained in Experiment 1 such that the samples lie inside the light cone with an upper temporal bound. Samples in the last row of Figure (a) had no constraints imposed on them. We observe larger morphological and location differences as time progresses. This is consistent with the theory that the system had enough time to evolve into these states. (b): In Experiment 2 we are intersecting 2 cones. For ease of reading the figures have been arranged such that the movements are more apparent. on the left in (b) we exhibit vertical movements while on the right we exhibit horizontal movements. The arrows guide the direction of reading in the figure.

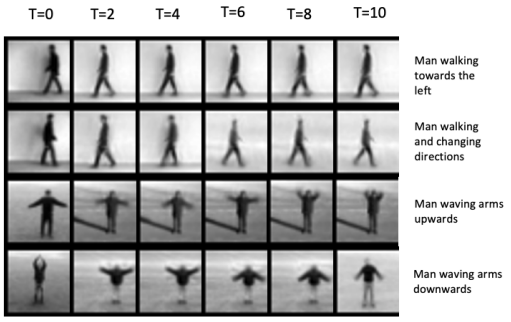
At a single time instant we can find any number of probable frames that extend a sequence. Hence, the `choose` step of Algorithm 1 depends on the target application. In this experiment to guide the choice of frames we map the sampled points to image space and compare the structural similarity of them with the original frame of $t = 0$. We adopt a simple manner to choose the next step and we do not provide the model with any further conditioning information to highlight the default strengths of the proposed algorithm. In an online inference scenario the reference frame could be updated as ground truth frames become available.

In Figures 2b and 3a we exhibit qualitative results of our algorithm when intersecting 2 and 5 cones respectively. In Figure 2b each set of results evaluates a specific movement, vertical or horizontal. In Figure 3a we exhibit the case of intersecting 5 cones. As this scenario allows up to 10 time steps for our model to evolve we notice a great number and more varied results. In the first two rows the depicted digits bounce while moving towards one direction. In the third row the digit 0 exhibits morphological changes and in the fourth row the digit 6 gradually moves its closing intersection upwards to become a 0. As our model is only trained with single frames of MNIST digits it is not constrained to show only movement or morphological changes. Rather it can vary both as seen in Figure 3a. The transmutation of the digit 6 to 0 is a probable, albeit unwanted, outcome under certain scenarios. In addition, we note that we are not providing any labels or additional information to the model during inference. In principle, one could condition the model to produce probable future frames by tuning the `choose` procedure of Algorithm 1. In the appendix we perform an analysis for the SSIM degradation over time and show how out method is not susceptible to autoregressive errors.

**Experiment 3: Realistic video data** As a final experimentation we use the KTH action dataset. Examples of the performance of the proposed algorithm are shown in Fig. 3b. Due to the computational constraints of the Poincaré VAE, which we are using as a base model, we are limited to one action at a time during training. We note how our algorithm retains characteristics like the shade of gray of the clothing while producing plausible futures. Each frame differs to the previous by 2 time instances giving ample time for the subject to change directions. We believe that with a higher capacity network a similar performance can be achieved on more complex scenes and higher resolution videos.




(a) Moving MNIST




(b) KTH movement video sequences dataset

Figure 3: Samples from Experiment 2 (a) and Experiment 3 (b). 5 cones intersected, trained on the moving MNIST dataset (a) and the KTH movement video dataset (b). Differences in image brightness in (b) are due to PyTorch's contrast normalization in the plotting function.

**Discussion** As the model is only trained as a VAE on single frames and not on sequences, the notion of time is not encoded in the weights of the network. Hence, all the resulting movement and predictive abilities are derived from our proposed algorithm and the natural embedding abilities of the Minkowski space-time.
We emphasize the time-agnostic nature of our algorithm. Our predictions are constrained in time but are probabilistic in nature. The proposed algorithm is able to produce multiple plausible futures. We believe this is a very important feature for future prediction and sequence extrapolation techniques as it can be used as an anomaly detection technique. Specifically, if one of the produced futures includes a hazardous situation, an automated system can adapt in order to avoid an outcome, enabling for example defensive driving capabilities in autonomous vehicles.
Even though our method is in principle auto-regressive, it does not suffer from the accumulation of errors as it is both probabilistic and relies on efficient latent space sampling rather than the ability of a neural network to remember structural and temporal information about the image.
Furthermore, we believe that the quality of the predicted frames as well as the definition of the subspace from which the samples should be derived could be improved by incorporating the inferred thermodynamic entropy of the frame. We will explore the link between the information and thermodynamic entropy in future work. In addition, even though our framework is architecture agnostic, a customized architecture for the prediction task would be an intriguing direction.
Finally, as our model allows us to find all probable scenarios that might exist, it can be used as a causal inference tool in the "potential outcomes" framework Rubin (2005). Given a state we are able to probe possible scenarios and investigate plausible outcomes, hence, deduce causal relations within the data. In addition, by using the *past* light cone $\mathcal{R}^-$, we are able to probe the events that could have led to an observed state enabling counterfactual analysis.

## 6 Conclusion

Machine Learning techniques are able to build powerful representations of large quantities of data. Leveraging this ability we propose that hard computer vision problems can be approached with minimal learning in an architecture agnostic manner. In this paper, we extend early Riemannian representation learning methods with the notion of Minkowski space-time as it is more suitable for causal inference. We further propose a novel algorithm to perform causally plausible image synthesis and future video frame prediction utilizing the special relativity concept of light cones and apply it to two standard datasets.

## REFERENCES

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic Entailment Cones for Learning Hierarchical Embeddings. *ICML*, 2018.

Dinesh Jayaraman, Frederik Ebert, Alexei Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *International Conference on Learning Representations*, 2019.

Olga Kosheleva and Vladik Kreinovich. Observable causality implies lorentz group: alexandrov-zeeman-type theorem for space-time regions. *Mathematical Structures and Modeling*, 2014.

Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart Russell, and Pieter Abbeel. Learning plannable representations with causal infogan. *Advances in Neural Information Processing Systems*, 2018.

Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection—A New Baseline. In *CVPR*, pp. 6536–6545. IEEE Computer Society, 2018.

Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders. *NeurIPS*, 2019.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, (Nips), 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. *35th International Conference on Machine Learning, ICML 2018*, 2018.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 2019.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics a primer*. Wiley, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference : Foundations and Learning Algorithms*. The MIT Press, 2019.

Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 2005.

Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pp. 32–36. IEEE, 2004.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. *ICML*, 2015.

Ke Sun, Jun Wang, Alexandros Kalousis, and Stephane Marchand-Maillet. Space-time local embeddings. In *NIPS*. 2015.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *CVPR*, pp. 1526–1535. IEEE Computer Society, 2018.

Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing Motion and Content for Natural Video Sequence Prediction. In *ICLR*, 2017a.

Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *ICML*, 2017b.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating Visual Representations from Unlabeled Video. In *CVPR*, pp. 98–106. IEEE Computer Society, 2016a.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *NIPS*, pp. 613–621, 2016b.

E. C. Zeeman. Causality implies the Lorentz group. *Journal of Mathematical Physics*, 5(4):490–493, 1964.