This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Spatio-temporal Predictive Network For Videos With Physical Properties

Yuka Aoyagi
Waseda University
yuka.aoyagi@akane.waseda.jp

Noboru Murata
Waseda University
noboru.murata@eb.waseda.ac.jp

Hidetomo Sakaino
Weathernews Inc.
sakain@wni.com

Abstract

In this paper, we propose a spatio-temporal predictive network with attention weighting of multiple physical Deep Learning (DL) models for videos with various physical properties. Previous approaches have been models with multiple branches for difference properties in videos, but the outputs of branches have been simply summed even with properties that change in time and space. In addition, it is difficult to train previous models for sufficient representations of physical properties in videos. Therefore, we propose the design of the spatio-temporal prediction network and the training method for videos with multiple physical properties, motivated by the Mixtures of Experts framework. Multiple spatio-temporal DL branches/experts for multiple physical properties and pixel-wise and expert-wise attention mechanism for adaptively integrating outputs of experts, i.e., Spatial-Temporal Gating Networks (STGNs) are proposed. Experts are trained with a vast amount of synthetic image sequences by physical equations and noise models. Instead, the whole network including STGNs is allowed to be trained only with a limited number of real datasets. Experiments on various videos, i.e., traffic, pedestrian, Dynamic Texture videos, and radar images, show the superiority of our proposed approach compared with previous approaches.

1. Introduction

Predicting future scenes has various industrial applications, such as automatic driving and anomaly detection in traffic scenes. Recently, various Deep Learning (DL) models with different architectures have been proposed to videos with traffic scenes [20] and on-board camera scenes [44]. However, predicted scenes and objects are suffered from distorted and blurry vehicles and pedestrians. In order to improve prediction accuracy, state-of-the-art (SOTA) DL models have been designed multiple branches for different properties in videos, e.g., background and moving objects [21]. In addition, physics-based DL models with branches [12] have been developed to follow underlying physical rules of videos. However, image intensity increments/decrements are not yet considered. Other SOTA multiple branches have been summed by static weights even under dynamic and local image property changes of input image sequences [17, 36, 42]. However, for designing a more expressive model, those weights should be determined adaptively based on input image sequences.

One of the promising approaches for adaptive weighting is to apply the framework of Mixtures of Experts (MoE). However, most MoE models consist of multiple non-spatio-temporal DL models/experts. Moreover, adaptive but scalar weightings to each expert output have been applied. Besides, in DL models of natural language processing and computer vision [38, 46], various attention mechanisms have been implemented to enhance feature representations of DL models, but no or less attention mechanism for video prediction has considered local and dynamic changes.

Training these complicated DL models generally requires a large amount of data [1], and end-to-end training with only real datasets is difficult. Training data augmentation by collection and annotation of real-world videos is important but expensive. Training methods of DL models on synthetic datasets have been proposed [15, 37], but they have not focused on the spatio-temporal dynamics of videos. Besides, since synthetic datasets are useful but have a wide deviation from real data, a certain bridge between them should be offered.

To this end, in this paper, we focus on designing a new spatio-temporal predictive network for videos with multiple physical properties, where DL models/experts and Spatio-Temporal Gating Networks (STGNs) with adaptive attention mechanism inspired by MoE are implemented. We call the proposed overall DL network, Spatio-Temporal MoE (STMoE). Contributions of this paper are four-fold:

1. Novel DL based prediction network, i.e., STMoE, is designed for video prediction, where an extended MoE consist of DL models/experts and STGNs, the former for extracting spatio-temporal feature representations of various local physical properties and the latter for determining contributions of multiple DL models.

2. STGNs are introduced to adaptively integrate expert-

derived feature representations, where a specific attention mechanism, i.e., a combination of pixel-wise attention for sub-regions of feature representations and expert-wise attention for physical property estimation, has been applied.

3. DL experts in STMoE are trained by synthetic image sequences from physical equations with different shapes, motion speeds, sizes, texture, translation, rotation, and image intensity increments. For filling the gap between real and synthetic images, a mix of natural phenomena driven noise, Perlin noise [14] and normal noise is proposed to efficiently augment training data.
4. A large amount of physics-based synthetic image sequences are used for training multiple experts. On the other hand, a less number of real image sequences are used to train the whole network. This new training framework ensures to eliminate costs to augment necessary real image data. Using many challenging scenes, i.e., dynamic MNIST, moving camera, moving legs, Dynamic Texture, and radar images, experimental results show improvements of proposed STMoE by predicted image quality over SOTA prediction models.

2. Related work

This section devotes Deep Learning (DL) based image and video prediction models and methods, dividing into four categories: non-physics, physics, multi-architecture, and attention mechanism models.

2.1. Non-physics based DL model

The basic principle of image prediction is to obtain the next single image frame from the past image sequences. In previous computer vision, the state space equations based 3D Auto-Regressive and Moving Average (ARMA) models have been used for Dynamic Texture (DT) videos, where DT is defined as time-varying and physics-rule-driven texture images/videos, i.e., traffic flow, pedestrian flow, natural phenomena [34]. Recently, in DL, ConvLSTM (Convolutional Long Short-Term Memory) [32] and its extended version, TrajGRU [33], PredRNN [41] and PredRNN++ [40] have been introduced. A variational inference model is also proposed [45]. More recently, ubiquitous U-Net [29] by collecting a large amount of training image datasets has been used to predict weather radar images [1]. Unlike collecting and annotating a large amount of real-world videos [7, 28, 48], this paper proposes to efficiently augment image sequences by physical equations, special noise, i.e., Perlin as well as different shapes, texture, and motion speeds.

2.2. Physics-based DL model

Exploiting prior physical knowledge is another appealing way to improve DL based prediction models. Among them, several approaches are dedicated to specific partial differential equations (PDEs). A specific architecture is designed for predicting and identifying a dynamical system [24]. PDE-Net [23] discretizes a broad class of PDEs by approximating partial derivatives with convolutions. Physics-based DL model, PhyDNet [12], is using a two-branch deep architecture, but there is no explicit modeling of intensity increments/decrements. Our proposed DL architecture has been designed to optimize importance of multiple physical branches/experts for a wider range of physical and natural phenomena in videos.

2.3. Multi-architecture DL model

In order to deal with different properties in videos, multiple DL architecture models have been proposed [36, 44], whose prediction accuracy is better than a single DL model. For the prediction of various motion features, a video prediction framework based on multi-frequency analysis has been proposed [20]. However, SOTA prediction models are hard to deal with strong local deformations and image intensity increments [17, 19, 21, 25, 26, 35] unlike our proposed local and physics based DL models.

2.4. Attention Mechanism

Attention mechanism dynamically determines important part of feature representations to pay attention to, during inference in DL depending on different inputs. In SOTA, multiple branches have been proven to be effective for multiple different tasks. There have been constant weightings of each branch [44]. Mixtures of Experts models [2, 3, 11, 16, 31, 39] have been used for scalar weightings or hard weightings, where multiple experts consist of classification based DLs. Squeeze and Excitation Network determining attention weights in each feature representation of multiple channels has achieved high accuracy in object recognition [18]. A number of papers to pixel-wise attention mechanisms for image scale [5] and for image/video context [8, 22, 47] have been succeeded to enhance accuracy. By following these avenues, our proposed Spatio-temporal Gating Networks (STGNs) play a role in pixel-wisely and expert-wisely weighting in response to input image sequences.

3. Proposed Methods

We propose a framework of designing Deep Learning (DL) model for video prediction, i.e., Spatio-Temporal Mixtures of Experts (STMoE). To begin with, we formulate a spatio-temporal prediction DL model. Let $Y_t \in \mathbb{R}^{M\times N\times C_Y}$ (M: height, N: width, C_Y: number of channels) be an image at time t and input data $X_t = \{Y_{t-1}, Y_{t-2}, ..., Y_{t-l}\}$ be temporally consecutive images at past time $t-1$ to $t-l$. In order to predict a next-time image Y_t from X_t, we generally define a spatio-temporal DL predictor $\mathcal{M}(\cdot\,;\theta)$:

$$\hat{Y}_t = \mathcal{M}(\{Y_{t-1}, Y_{t-2}, ..., Y_{t-l}\}\,;\theta) = \mathcal{M}(X_t\,;\theta), \quad (1)$$

where θ is a parameter to be trained. For longer frame prediction, an updating output image is recursively used to a new input image until future time $t+n-1$. Suppose that $\mathcal{M}(\cdot\,;\theta)$ consists of an encoder $F(\cdot\,;\theta_F)$ and a decoder $H(\cdot;\theta_H)$, where $F(\cdot\,;\theta_F)$ obtains a feature representation $Z_t \in \mathbb{R}^{P\times Q\times C_Z}$($P$: height, Q: width, C_Z: number of channels) from X_t and $H(\cdot;\theta_H)$ obtains $\hat{Y}_t$ from Z_t.

$$Z_t = F(X_t\,;\theta_F), \quad (2)$$

$$\hat{Y}_t = H\big(Z_t\,;\theta_H\big) = H\big(F(X_t\,;\theta_F);\theta_H\big). \quad (3)$$

Since in this paper, multiple local and physical properties, e.g., rotation and intensity increments, are assumed, and Z_t is approximated by the linear combination of different m distinct physical properties:

$$Z_t = \sum_{k=1}^{m} Z_t^k. \quad (4)$$

3.1. Spatio-Temporal Mixtures of Experts

In this section, methods for obtaining feature representations of corresponding physical properties and the integration of them as in Equation 4, are described in the framework of proposed STMoE. STMoE consists of m distinct spatio-temporal DL physical encoders/experts $\{F^k(\cdot\,;\theta_{F^k}), k = 1, ..., m\}$ and Spatio-temporal Gating Networks (STGNs), $\{G^k(\cdot\,;\theta_{G^k}), k = 1, ..., m\}$ as shown in Figure 1(a). The physical feature representation $\tilde{Z}_t^k \in \mathbb{R}^{P\times Q\times C_Z}$ corresponding to the the kth physical property is estimated by $F^k(\cdot\,;\theta_{F^k})$ trained with specific physical dynamics. However, it is difficult for $F^k(\cdot\,;\theta_{F^k})$ to extract only the feature representation necessary for the kth property from X_t which includes multiple spatio-temporally varying physical properties, resulting in $\tilde{Z}_t^k$ containing spatial sub-regions which don't respond to the kth property. Therefore, a pixel-wise attention weight $W_t^k \in \mathbb{R}^{P\times Q}$ where $Z_t^k = W_t^k \cdot \tilde{Z}_t^k$, $0 \le W_t^k \le 1$, is needed to ensure that only spatial regions that respond to the kth property are extracted. The design of attention weights $W_t = \{W_t^k, k = 1, ..., m\}$ is shown in Figure 1(b). W_t are calculated based on past spatio-temporal changes around the location to be weighted. First, for attention weights, a new feature representation $U_t \in \mathbb{R}^{P\times Q\times m'}$ (m': number of channels) with the same height and width of Z_t is caluculated from $G'(\cdot\,;\theta_{G'})$, the first part of STGNs where input is X_t, i.e., $U_t = G'(X_t\,;\theta_{G'})$. The feature vector at the pixel of position i, j in the physical feature representation obtained from $F^k(\cdot\,;\theta_{F^k})$ is denoted as $\tilde{Z}^k[i,j]$, and the corresponding scalar attention weight is denoted as $W_t^k[i,j]$. Then the kth attention weight of pixel i, j, i.e., $W_t^k[i,j]$ are calculated as follows:

$$\begin{aligned} W_t^k[i,j] &= G^{k''}\big(U_t[R(i,j)];\theta_{G^{k''}}\big) \\ &= G^k(X_t;\ \theta_{G^k})[i,j], \end{aligned} \quad (5)$$

where spatial regions around pixel i, j is denoted as $R(i,j)$ and $G^{k''}(\cdot\,;\theta_{G^{k''}})$ is the second part of STGN. The function of proposed STGNs can be thought as a specific attention mechanism; a combination of pixel-wise attention for sub-regions of feature representation and expert-wise attention for physical property estimation. Furthermore, since physical features are assumed to be similar in neighboring pixels in Z_t, spatial smoothness has been posed to W_t. We use convolutional neural network and squeeze and excitation network[18] for STGNs for spatial smoothness of W_t and spatial constraints with $R(i,j)$. From the above, a predicted image at time t, i.e., $\hat{Y}_t$ is given in Equation 6:

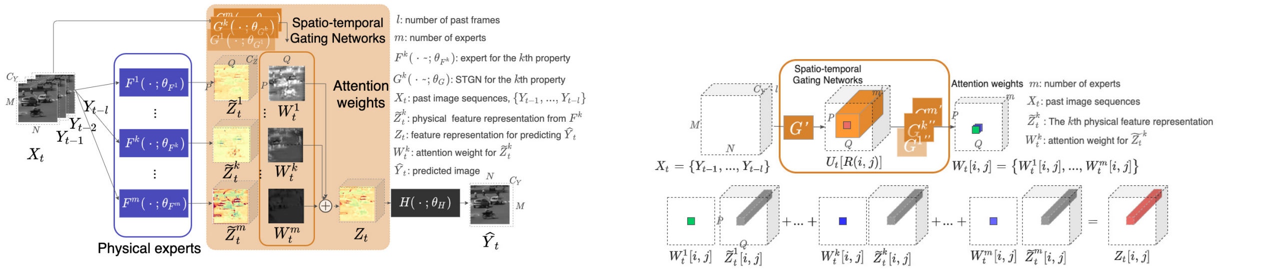
$$\begin{aligned} \hat{Y}_t &= H\big(\sum_{k=1}^{m} W_t^k \cdot \tilde{Z}_t^k\,;\theta_H\big) \\ &= H\big(\sum_{k=1}^{m} G^k(X_t\,;\theta_{G^k}) \cdot F^k(X_t\,;\theta_{F^k})\,;\theta_H\big). \end{aligned} \quad (6)$$

3.2. Optimization of STMoE

A two-step optimization for STMoE is explained: local optimization, which is mainly for training of DL experts, and global optimization, which is for training of the whole, including STGNs.

Local optimization: We propose a method for training multiple experts using newly augmented synthetic image sequences and a training process for constructing feature representations that can be integrated by linear combination as shown in Equation 6. Each of all experts $\{F^k(\cdot\,;\theta_{F^k}), k = 1, ..., m\}$ needs to be trained independently to extract the corresponding feature representation. Due to insufficient training datasets, synthetic image sequences $\mathcal{D}^k$ for the kth property is used for training $F^k(\cdot\,;\theta_{F^k})$. In order to facilitate synthesizing a vast amount of $\mathcal{D}^k$ efficiently, simple but physical equations have been employed. Moreover, since this paper aims at predicting complicated videos capturing natural phenomena with mixed physical properties, various noises i.e., Gaussian, White, and Perlin noise [14], have been added to make experts robust to real videos. In particular, Perlin noise is known as a model of natural phenomena. An example of generation of synthetic image sequences is shown in Figure 2. Details are as follows:

- Initial objects are placed in a frame. Objects are generated from circles, lines, polygons. Initial textures are generated from a sine wave.



(a) Overall network. STMoE consists of multiple experts corresponding to physical properties and STGNs which determine attention weights.

(b) Pixel-wise attention mechanism for physical feature representations.

Figure 1. Proposed Spatio-Temporal MoE (STMoE): (a) shows the overall network, and (b) shows details of the designed attention mechanism in the orange shaded area in (a).

- Their motions and image intensity changes over time as:
  - Translation, rotation: $\mathcal{D}^k$ is generated by changing objects' velocity/angular velocity, acceleration/angular acceleration and jerk/angular jerk.
  - Intensity increments/decrements: $\mathcal{D}^k$ is generated by applying convolution filters to images at each time step. When the total filter value is higher/lower than 1, it simulates increments/decrements, respectively.

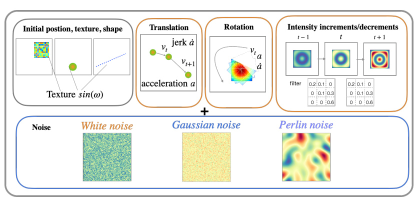


Figure 2. An example of how to generate synthetic image sequences $\{\mathcal{D}^k, k = 1, ..., m\}$.

In order to obtain feature representations which can be linearly combined, a two-stage training process has been implemented as follows:

1. Training of $H(\cdot\,;\theta_H)$: A certain spatio-temporal DL encoder $F^0(X_t\,;\theta_{F^0})$, e.g., MIM is prepared for training of $H(\cdot\,;\theta_H)$. By setting $\hat{Y}_t = H(F^0(X_t\,;\theta_{F^0})\,;\theta_H)$, parameters θ_H and θ_{F^0} are trained using synthetic image sequences including all physical properties, $\mathcal{D}^{all} = \{\mathcal{D}^k, k = 1..., m\}$. Generalized model parameters θ_H and θ_{F^0} for all physical properties can be obtained. Only θ_H is used in the following training process.
2. Training of all experts $\{F^k(\cdot\,;\theta_{F^k}), k = 1, ..., m\}$: Training of $F^k(\cdot\,;\theta_{F^k})$ is conducted by using $\mathcal{D}^k$. When training any expert, a shared θ_H is fixed, so that all experts obtain linearly combinable feature representations.

The loss function of local optimization follows the usual loss function of DL experts. Each of experts can be trained independently and simultaneously, therefore the time cost is not expensive.

Global optimization: After local optimization, the whole network including STGNs is trained with a limited number of real datasets. To optimize the whole network, we define a loss function which consists of three objective: 7:

$$\mathcal{L} = \mathcal{L}_{reconst} + \lambda_{Ws} \cdot \text{smo}(W) + \lambda_{We} \cdot \text{ent}(W). \tag{7}$$

The first term of Equation 7 is reconstruction error $\mathcal{L}_{reconst} = \|\hat{Y} - Y\|$, $Y = \{Y_t, ..., Y_{t+n-1}\}$. The second and third term of Equation 7 are constraints on attention weights $W = \{W_t, W_{t+1}, ..., W_{t+n-1}\}$. In video predictions, W needs smoothly varying in the temporal direction, the second term, L^1 constraint of consistency in the temporal direction $\text{smo}(W) = \|\frac{\partial}{\partial t}W\|_1$ is introduced. Moreover, the third term, $\text{ent}(W) = -W \log W$, is introduced for promoting *sparsity* of W, i.e., extracting as few important features as possible. In addtion, to deal with outliers and limited datasets, a generalized robust function $\rho(x, \alpha, c)$ is employed [4]. Thus, the final objective function is:

$$\begin{aligned}\mathcal{L} = {} & \rho(\mathcal{L}_{reconst}, \alpha_1, c_1) + \\ & \lambda_{We} \cdot \rho(\text{ent}(W), \alpha_2, c_2) + \lambda_{Wm} \cdot \rho(\text{mag}(W), \alpha_3, c_3).\end{aligned} \tag{8}$$

Global optimization is done by minimizing Equation 8 with respect to parameters $\theta_F = \{\theta_{F^k}, k = 1, ..., m\}, \theta_G = \{\theta_{G^k}, k = 1, ..., m\}$, and θ_H. For ensuring convergence, global optimization has been divided into two stages:

1. Training of $G(\cdot\,;\theta_G)$ is conducted by fixing $\theta_F = \{\theta_{F^1}, ..., \theta_{F^m}\}$ and θ_H.
2. Training of the whole network is conducted and trained θ_F, θ_G, and θ_H are obtained.

4. Experiments

We have conducted experiments to justify proposed Spatio-temporal Mixtures of Experts (STMoE) by comparing with five state-of-the art (SOTA) baseline models: U-Net [1], Advection Diffusion into DL (ADDL) [6], PredRNN++ [40], Memory in Memory (MIM) [43] and PhyDNet [12]. As examples of STMoE, following models were used. Since video are assumed to be composed of multiple physical properties as described in Section 3, three basic physical properties ($m = 3$), i.e., translation, rotation, and intensity increments/decrements were used. The other physical properties in videos were assumed to be represented approximately by a combination of these properties. Two models with different DL experts and decoder $H((\cdot\,;\theta_H))$, i.e., STMoE-Id and STMoE-conv, were used.

STMoE-Id: In STMoE-Id, an identity decoder was used for $H(\cdot\,;\theta_H)$, i.e., $Z_t = \hat{Y}_t$. A model excelled in prediction accuracy of $\mathcal{D}^k$ was selected for $F^k(\cdot\,;\theta_{F^k})$. The model that incorporates physical structures of ADDL [6] was used for the expert of translation. For other experts, MIM was used [43], where MIM takes a temporal difference in hidden states and is useful for non-stationary components.

STMoE-conv: In STMoE-conv, convolution filter was used for $H(\cdot\,;\theta_H)$ in Equation 6. By using convolution filter for $H(\cdot\,;\theta_H)$, it was assumed that spatial distribution of physical properties would be smooth in $\hat{Y}_t$. In STMoE-conv, MIM was used for all experts.

Since ADDL and MIM are experts of STMoE, they were used to compare prediction accuracy between ADDL/MIM and STMoE. Experiments were conducted using both synthetic data, i.e., Dynamic MNIST, and real data, i.e., traffic scenes, pedestrian scenes, Dynamic Texture (DT) videos, and precipitation radar images in Figure 3. All data were downsampled and cropped to $M, N = 112$ or 128 and converted to grayscale ($C_Y = 1$) due to computational resources. Synthetic image sequences $\mathcal{D}^{all}$ for local optimization of STMoE-Id and -conv were generated totally 30,000 sequences (10,000 sequences per expert). 4 temporally consecutive image frames ($l = 4$) are used in order to estimate velocity, acceleration and jerk of motions. Other hyperparameters were determined based on validation datasets. All baseline models were trained using only real datasets. Quantitative evaluations are carried out using two metrics: Structural Similarity (SSIM) to analyze local structures of videos and Mean Square Error (MSE) to evaluate image intensities of each pixel especially for precipitation radar images. Note that minimum MSE (↓) and maximum SSIM (↑) indicate the best performance in tables.

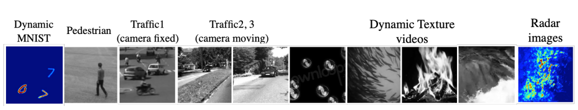


Figure 3. Datasets used in experiments.

models	MSE ↓	SSIM ↑	models	MSE ↓	SSIM ↑
U-Net [1]	336	0.920	ADDL [6]	213	0.964
PredRNN++ [40]	205	0.933	PhyDNet [12]	162	0.928
MIM [43]	200	0.963	MIM w/$\mathcal{D}^{all}$[43]	174	0.964
STMoE-Id	**115**	**0.978**	STMoE-conv	**120**	**0.975**

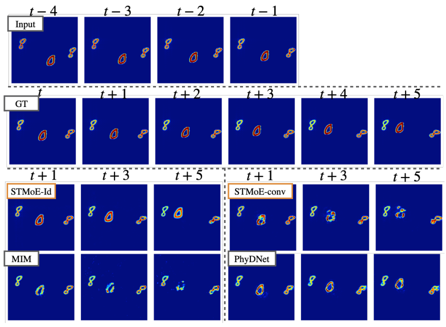
Table 1. Quantitative results on Dynamic MNIST.

4.1. Dynamic MNIST

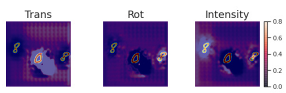
Dynamic MNIST extended Moving MNIST [32] with translation, rotation, and intensity change was used since Moving MNIST was not enough for evaluating multiple physical properties. 6 frames were predicted. The number of training data was 250. Table 1 shows proposed STMoE-Id and -conv outperform baseline models. In addition to baseline models trained only with Dynamic MNIST, MIM trained with both Dynamic MNIST and synthetic image sequences $\mathcal{D}^{all}$ used in local optimization of STMoE (w/$\mathcal{D}^{all}$: with $\mathcal{D}^{all}$) was also examined to eliminate differences in the number of total training datasets. Input 4-frame images, ground truth (GT) images, and prediction images of MIM, PhyDNet, and STMoE model are shown in Figure 4(a), where the middle '0' is translating, the right '8' is rotating, and the intensity of the left '8' is gradually decreasing. Prediction results of STMoE-Id and -conv are well predicted physical properties compared to prediction results of MIM and PyhDNet, as blur and intensity degradation have been improved. In order to better understand roles of three physical experts and Spatio-temporal Gating Networks (STGNs) with pixel-wise attention weighting, Figures 4(b) and 4(c) show predicted results overlaid overlaid with importance of each experts for prediction, i.e., attention weights $\{H(W_t^k\,;\theta_H), k = 1, 2, 3\}$, where highlighted spatial regions indicate high importance of corresponding experts. In Figures 4(b) and 4(c) , the middle '0', the right '8', and the left '8' have high importance of translation expert (Trans), rotation expert (Rot), and intensity increments/decrements expert (Intensity), respectively. It has been confirmed that proposed STGNs have functioned as designed.

4.2. Pedestrian and Traffic scenes

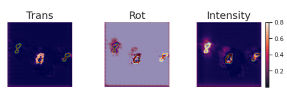
In this section, experiments were conducted using four scenes: pedestrian scenes from KTH [30] (Pedestrian), traffic scenes with pedestrians and vehicles from the public dataset DT videos [13] (Traffic 1), and traffic scenes with the on-board camera from KITTI [10, 27] (Traffic2, 3). 6 frames or 4 frames were predicted. STMoE-Id was used for prediction. The number of training data was from a few hundred to 1500. Table 2 shows that STMoE-Id has the best performance among all in terms of SSIM. Figure 5(a) shows that STMoE-Id is able to capture cars and pedestrians well on Traffic1 scene. Enlarged car regions show that the blurring and distortion of car in the baseline has been improved when STMoE-Id is applied. For better analyzing lo-



(a) Comparative results of Dynamic MNIST prediction.



(b) Attention weights in STMoE-Id

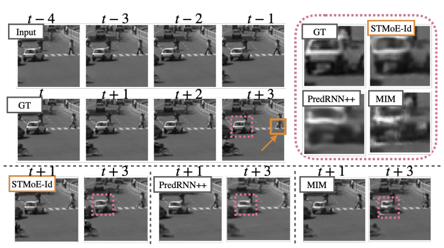


(c) Attention weights in STMoE-conv

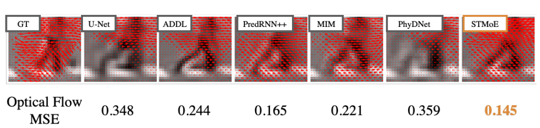
Figure 4. Experiment on Dynamic MNIST.

models	Pedestrian ↑	Traffic 1 ↑	Traffic 2 ↑	Traffic 3 ↑
U-Net [1]	0.807	0.867	0.375	0.364
ADDL [6]	0.872	0.904	0.422	0.386
PredRNN++ [40]	0.875	0.890	0.452	0.402
MIM [43]	0.865	0.890	0.431	0.405
PhyDNet [12]	0.827	0.846	0.461	0.419
STMoE-Id	**0.877**	**0.909**	**0.467**	**0.427**

Table 2. Evaluation by SSIM on pedestrian and traffic scenes.



(a) Predicted images of STMoE-Id, PredRNN++ and MIM. Pink dotted regions show enlarged images around the car.



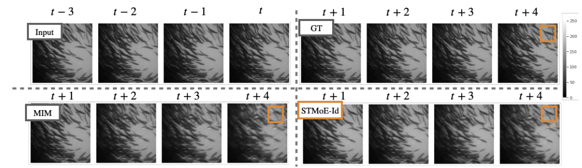
(b) Evaluation of pedestrian's legs' motions by Optical Flow in orage rectangular region of (a).

Figure 5. Experiment on Traffic1 scene.

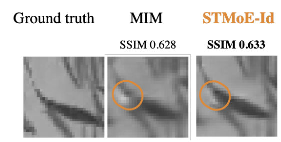
cal legs' motions of the pedestrian, we applied Optical Flow (OF) [9] to prediction results and used MSE of OF in Figure 5(b). As shown quantitatively by local MSE of OF, whereas baseline models do not capture legs moving downward to the right, STMoE-Id captures them. It has been confirmed that STMoE-Id, which takes multiple physical features into account pixel-wisely, is better able to capture local motion and intensity changes than baselines.

	Bubble ↑	Fish ↑	Fire ↑	River ↑
U-Net [1]	0.859	0.701	0.414	0.715
ADDL [6]	0.903	0.782	0.488	0.697
PredRNN++ [40]	0.950	0.854	0.402	**0.745**
MIM [43]	0.956	0.858	0.432	0.741
PhyDNet [12]	0.834	0.761	0.411	0.729
STMoE	**0.970**	**0.863**	**0.489**	0.741

Table 3. Quantitative results on DT videos by SSIM. Note that better results in STMoE-Id or -conv are shown.



(a) Predicted images.



(b) Local evaluation and enlarged images of orange rectangular regions in (a).

Figure 6. Comparative results of STMoE-Id and MIM prediction using Fish.

4.3. Dynamic Texture Videos

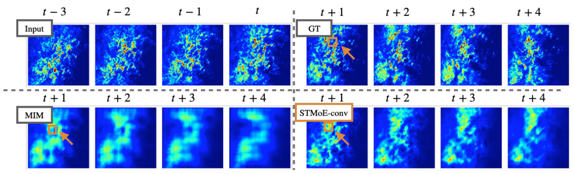
Dynamic texture is a scene where objects with texture changes spatio-temporally in response to physical phenomena, four challenging videos were used from DT videos [13]: Bubble (semi-transparency, elastic body), Fish (school, elastic body), Fire (semi-transparency, fluidity), and River (rough wave, fluidity). The number of training data was from 200 to 1,000. Table 3 presents STMoE superior to baseline models except for River. Note that better results in STMoE-Id or -conv are shown in Table 3. Figure 6(a) shows prediction images of MIM and STMoE-Id on Fish and Figure 6(b) presents magnified images of orange rectangular regions of Figure 6(a). Image quality degradation such as white extra spots around fish is observed in MIM's prediction results; however, such image degradation is reduced in proposed STMoE's prediction results. SSIM in this local region supports these results.

4.4. Precipitation Radar Images

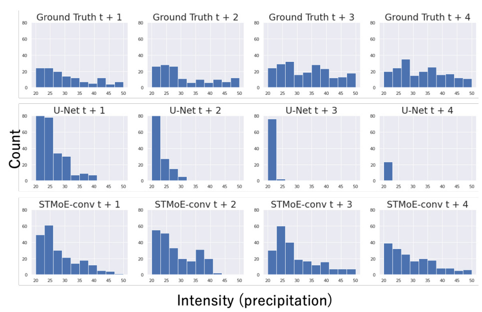
In this section, precipitation radar images have been used as a challenging example of DT with intensity increments/decrements, i.e., growth/decay of precipitation intensity. Furthermore, the frame rate is 10 minutes, which is longer than that of other videos. The high intensity of precipitation radar images is important because it corresponds to heavy rainfall, which can lead to major disasters. 4 frames were predicted. The number of training data was

models	MSE ↓	B-MSE/10 ↓
U-Net [1]	53.0	112
ADDL [6]	58.3	117
PredRNN++ [40]	58.0	115
MIM [43]	57.9	117
PhyDNet [12]	59.7	120
STMoE-conv	**51.6**	**103**

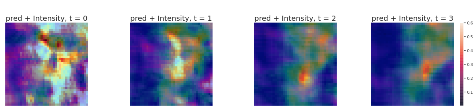
Table 4. Quantitative results on precipitation radar images.



(a) Predicted image results.



(b) Temporal variation of the histogram of image intensity, i.e., precipitation intensity, in orange rectangular regions in (a).



(c) Attention weights of intensity increments/decrements experts in the orange rectangular region in (a).

Figure 7. Comparative results of STMoE-conv and U-Net prediction using precipitation radar images.

about 9,000 samples. STMoE-conv was used to predict. Balanced-MSE (B-MSE) is added to evaluate heavy rainfall used in [32]. Table 4 shows that STMoE-conv outperforms in terms of prediction error in two metrics. In prediction results of Figure 7, locally high image intensity is successfully predicted by STMoE-conv, whereas blurred image without high image intensity is predicted by U-Net. Figure 7(b) shows the temporal changes of distribution of image intensities, i.e. precipitation intensities, in extracted orange rectangular regions of Figure 7(a). Ground truth and proposed STMoE-conv show an increase of image intensities in the range of 30 to 50, whereas the U-Net does not capture it. As can be seen in Figure 7(c), attention weights of intensity increments/decrements experts are applied in the orange rectangular region, indicating that intensity changes are captured by proposed STMoE.

methods	Pedestrian		Precipitation	
	MSE	SSIM	MSE	B-MSE/10
STMoE w/local and global optimization	**59.6**	**0.877**	**51.6**	**103**
STMoE w/end-to-end	61.3	0.873	52.7	109

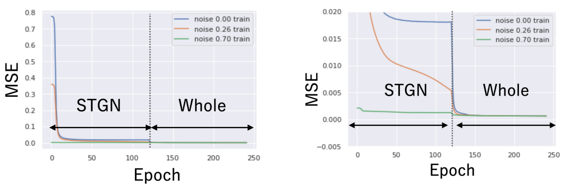
Table 5. Quantitative evaluation of two different optimization methods on pedestrian scene and precipitation radar images.

4.5. Ablation Studies

We provide detailed analysis of four contributions of proposed STMoE: 1) training expert with physical synthetic image sequences, 2) adding noise to synthetic image sequences, 3) pixel-wise attention weighting of expert-derived physical feature representations, and 4) training on a limited number of real datasets.

Training Expert with Physical Synthetic Image Sequences: In order to verify the effectiveness of training of multiple experts with synthetic images sequences, i.e., local optimization, experiments were performed on Pedestrian scene and precipitation radar images with two methods: STMoE with local and global optimization (STMoE architecture with local and global optimization, proposed STMoE), and STMoE with end-to-end optimization (STMoE architecture with end-to-end optimization using only real datasets). Table 5 shows that the prediction accuracy of STMoE with local and global optimization is better than that of STMoE with end-to-end. It has been suggested that STMoE with training of multiple experts with physical synthetic image sequences is effective to improve prediction accuracy and to solve the difficulty of training DL prediction models using only real datasets.

Adding Noise to Synthetic Image Sequences: This paper proposes to add noise to synthetic image sequences when training experts in local optimization in order to obtain robust experts to real datasets. Experiments for this were conducted with precipitation radar images. STMoE-conv was used and the number of training datasets was about 800. We prepared three types of data with ratio of the total intensity of the added mixed noises (Gaussian, White, and Perlin noise) to that of the synthetic image sequences: $\{0, 0.26, 0.70\}$. As a result, B-MSE was $\{1.05 \times 10^3, 1.02 \times 10^3, 1.00 \times 10^3\}$ for noise ratios $\{0.00, 0.26, 0.70\}$ respectively. The highest noise ratio 0.70 stands for the best prediction accuracy. It has been suggested that additive noise can enhance the prediction accuracy of real datasets. Moreover, for better understanding of the effect of additive noise, the convergence in training is taken into account. Figure 8 shows results by comparing convergence with three noise levels. Also, in a two-stage global optimization of STGNs and the whole network, the fastest convergence is obtained when noise-0.70 in green and lowest when noise-0.00 in blue, showing over 10 times



(a) MSE during training (b) Enlarged version of (a)

Figure 8. Comparison of convergence in training STGNs and the whole network with three noise levels to synthetic image data.

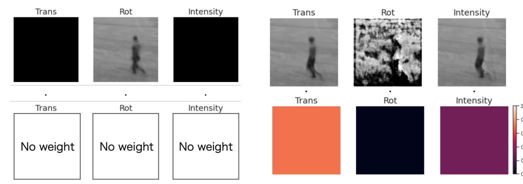
Attention methods	MSE ↓	SSIM ↑
No attention	75.0	0.864
Uniform attention	70.2	0.870
Pixel-wise attention	**59.6**	**0.877**

Table 6. Quantitative results with three cases of attention.

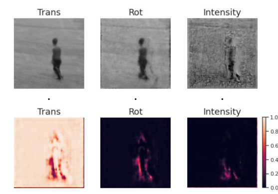
improvement. These results have suggested that robust experts generalized to real datasets can be obtained by training experts on synthetic image sequences with mixed noise in local optimization.

Pixel-Wise Attention of Expert-derived Physical Feature Representations: In order to verify the effects of pixel-wise attention to experts-derived feature representations in STMoE, experiments have been conducted with three attention cases to STMoE-Id: with no attention to experts (i.e., no gating networks), with uniform attention (i.e., conventional MoE), and with pixel-wise attention (i.e., proposed STMoE). Pedestrian datasets were used. Table 6 shows that pixel-wise attention case is the best result over no attention and uniform attention cases. Figure 9 shows the visualization of experts' outputs and attention weights in the space of Y_t, i.e., $\{H(\tilde{Z}_t^k; \theta_H), k = 1,2,3\}$ and $\{H(W_t^k; \theta_H), k = 1,2,3\}$ in experiment using Pedestrian scenes. Figure 9(a) shows no attention case, where only one expert is used. Figure 9(b) shows uniform attention case, where two experts are used with scalar weights. Compared to the above two attention cases, pixel-wise attention case in Figure 9(c) is the most expressive, where translation expert is used for the body part, and rotation and intensity increments/decrements experts are used for legs. Thus, proposed pixel-wise attention mechanism has been shown to be effective in predicting local physical properties.

The Number of Real Datasets in Training: Experiments were conducted to verify the prediction accuracy of proposed STMoE with a small number of training datasets on Traffic 1. Proposed STMoE-Id, MIM, and MIM with $\mathcal{D}^{all}$ whose prediction results are second best in Table 2 were used. Figure 10 shows SSIM when the number of real videos used for training is varied. STMoE shows less degradation in SSIM with decreasing the number of real videos in training than MIM and MIM with $\mathcal{D}^{all}$. In



(a) No attention (b) Uniform attention



(c) Pixel-wise attention

Figure 9. Visualization of decoded images of three experts, i.e., $\{H(\tilde{Z}_t^k; \theta_H), k = 1,2,3\}$ (upper row) and attention weights, i.e., $\{H(W_t^k; \theta_H), k = 1,2,3\}$ (bottom row) on Pedestrian experiment.

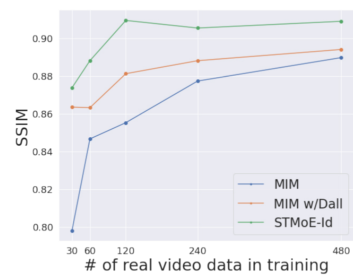


Figure 10. SSIM when using different number of real-datasets for training among three prediction models.

addition, SSIM of STMoE is better than that of MIM with $\mathcal{D}^{all}$, indicating that STMoE is more effective with limited real videos in training than other baseline models even when the same synthetic image sequences are used.

5. Conclusion

This paper has proposed a adaptive network with attention weighting to multiple physical Deep Learning models/experts, i.e., Spatio-Temporal Mixtures of Experts (STMoE), for real-world video prediction. Image sequences with multiple local and dynamic time-varying physical properties are pixel-wisely and expert-wisely taken into account by using proposed Spatio-Temporal Gating Networks, whereas SOTA DL models have used static or scalar weightings to branches with different experts. Experimental results on both synthetic and real data have shown the superiority of our STMoE in comparison with SOTA approaches in terms of less blur and distortion, in particular, representations of local dynamics. In future work, more complicated video scenes under time-varying illumination or with periodic motions will be addressed.

References

[1] Shreya Agrawal, Luke Barrington, Carla Bromberg, John Burge, Cenk Gazen, and Jason Hickey. Machine learning for precipitation nowcasting from radar images. *CoRR*, abs/1912.12132, 2019. 1, 2, 5, 6, 7

[2] Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9911 LNCS:516–532, 2016. 2

[3] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7120–7129. IEEE Computer Society, 2017. 2

[4] J. T. Barron. A general and adaptive robust loss function. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4326–4334, June 2019. 4

[5] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3640–3649. IEEE Computer Society, 2016. 2

[6] Emmanuel de Bézenac, Arthur Pajot, and Patrick Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 5, 6, 7

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[8] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8554–8564. Computer Vision Foundation / IEEE, 2019. 2

[9] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Josef Bigün and Tomas Gustavsson, editors, *Image Analysis, 13th Scandinavian Conference, SCIA 2003, Halmstad, Sweden, June 29 - July 2, 2003, Proceedings*, volume 2749 of *Lecture Notes in Computer Science*, pages 363–370. Springer, 2003. 6

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[11] Sam Gross, Marc'Aurelio Ranzato, and Arthur Szlam. Hard mixtures of experts for large scale weakly supervised vision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5085–5093. IEEE Computer Society, 2017. 2

[12] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5, 6, 7

[13] Isma Hadji and Richard P. Wildes. A new large scale dynamic texture dataset with application to convnet understanding. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11218 LNCS:334–351, 2018. 5, 6

[14] John C. Hart. Perlin noise pixel shaders. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Workshop on Graphics Hardware*, HWWS '01, page 87–94, New York, NY, USA, 2001. Association for Computing Machinery. 2, 3

[15] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. *CoRR*, abs/1710.10710, 2017. 1

[16] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 2

[17] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Fei-Fei Li, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 515–524, 2018. 1, 2

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 2, 3

[19] Yue Hu, Siheng Chen, Ya Zhang, and Xiao Gu. Collaborative motion prediction via neural motion message passing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[20] Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, and Xiaowei Li. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[21] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: Multiple frames prediction for learned video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[22] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3089–3098. IEEE Computer Society, 2018. 2

[23] Zichao Long, Yiping Lu, and Bin Dong. Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399, Dec. 2019. 2

[24] Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In Jennifer G.

Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3282–3291. PMLR, 2018. 2

[25] Osama Makansi, Ozgun Cicek, Kevin Buchicchio, and Thomas Brox. Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[26] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Mantra: Memory augmented networks for multiple trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[27] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[28] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *CoRR*, abs/1801.03150, 2018. 2

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 2

[30] Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004*, pages 32–36. IEEE Computer Society, 2004. 5

[31] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–19, 2017. 2

[32] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810, 2015. 2, 5, 7

[33] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and A new model. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5617–5627, 2017. 2

[34] S. Soatto, G. Doretto, and Y. N. Wu. Dynamic textures. *Proceedings of the IEEE International Conference on Computer Vision*, 2(C):439–446, 2001. 2

[35] Hao Sun, Zhiqun Zhao, and Zhihai He. Reciprocal learning networks for human trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[36] Matthew Tesfaldet, Marcus A. Brubaker, and Konstantinos G. Derpanis. Two-stream convolutional networks for dynamic texture synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[37] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 969–977. IEEE Computer Society, 2018. 1

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1

[39] Xin Wang, Jiawei Wu, Da Zhang, Yu Su, and William Yang Wang. Learning to compose topic-aware mixture of experts for zero-shot video captioning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 8965–8972, 2019. 2

[40] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Predrnn++: Towards A resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5110–5119. PMLR, 2018. 2, 5, 6, 7

[41] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):880–889, 2017. 2

[42] Yunbo Wang, Jiajun Wu, Mingsheng Long, and Joshua B. Tenenbaum. Probabilistic video prediction from noisy data with a posterior confidence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[43] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:9146–9154, 2019. 5, 6, 7

[44] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[45] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, and Trevor Darrell. Video prediction via example guidance. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10628–10637. PMLR, 13–18 Jul 2020. 2

[46] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10073–10082. IEEE, 2020. 1

[47] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13062–13071. IEEE, 2020. 2

[48] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 2