S-HR-VQVAE: Sequential Hierarchical Residual Learning Vector Quantized Variational Autoencoder for Video Prediction

Mohammad Adiban[†], Kalin Stefanov[*], Sabato Marco Siniscalchi[†+] *Senior Member, IEEE,* Giampiero Salvi[†] *Senior Member, IEEE*

[†]Norwegian University of Science and Technology
[*]Monash University
[+]Università degli Studi di Palermo
E-mails: {mohammad.adiban,marco.siniscalchi,giampiero.salvi}@ntnu.no, kalin.stefanov@monash.edu

***Abstract*—We address the video prediction task by putting forth a novel model that combines (i) a novel hierarchical residual learning vector quantized variational autoencoder (HR-VQVAE), and (ii) a novel autoregressive spatiotemporal predictive model (AST-PM). We refer to this approach as a sequential hierarchical residual learning vector quantized variational autoencoder (S-HR-VQVAE). By leveraging the intrinsic capabilities of HR-VQVAE at modeling still images with a parsimonious representation, combined with the AST-PM's ability to handle spatiotemporal information, S-HR-VQVAE can better deal with major challenges in video prediction. These include learning spatiotemporal information, handling high dimensional data, combating blurry prediction, and implicit modeling of physical characteristics. Extensive experimental results on four challenging tasks, namely KTH Human Action, TrafficBJ, Human3.6M, and Kitti, demonstrate that our model compares favorably against state-of-the-art video prediction techniques both in quantitative and qualitative evaluations despite a much smaller model size. Finally, we boost S-HR-VQVAE by proposing a novel training method to jointly estimate the HR-VQVAE and AST-PM parameters.**

***Index Terms*—Video Prediction, Hierarchical Modeling, Autoregressive Modeling**

I. Introduction

Video prediction involves anticipating future video frames based on a sequence of preceding frames [1]. It is a challenging task, requiring algorithms to grasp complex spatiotemporal relationships within the video, at the same time as handling high dimensionality, addressing blurry predictions, and accounting for the physical characteristics of the scenes. *Spatiotemporal modeling* aims to capture dependencies in video frame sequences, mirroring human perception of dynamic phenomena [2]. This is a general problem in video modeling, but becomes especially challenging when we need to recursively and accurately predict video frames for long temporal spans. Current state-of-the-art methods often struggle with long-term dependencies and complex motion patterns, leading to inaccuracies in the predicted frames. *High dimensionality* is inherent in video patterns, leading to the “curse of dimensionality” in function approximation and optimization [3]. Autoencoder-based methods attempt to reduce dimensionality, but may lose important fine-grained details necessary for accurate prediction. *Blurry predictions* stem from statistical models producing fuzzier outputs when predicting uncertain future events. This is, therefore, a more challenging problem for video prediction than for any other video task. Most methods use mean squared error (MSE) objective that tends to average over possible outcomes, resulting in blurred predictions. The challenge of *physical characteristics* pertains to object and scene attributes affecting prediction. Proper modeling of these characteristics may potentially aid future frame predictions. Recent video prediction methods have made significant progress in tackling these challenges, yet they still face several limitations. We will detail the state-of-the-art with respect to each of these challenges in Section II.

This paper introduces a sequential hierarchical residual learning vector quantized variational autoencoder (S-HR-VQVAE), which is tailored for video prediction with the goal of tackling the above-mentioned challenges. To this end, S-HR-VQVAE implements a novel autoregressive spatiotemporal predictive model (AST-PM) to capture distributions of dependencies between latent representations across time and space. The latent representations are generated through our novel encoding scheme, termed hierarchical vector quantization variational autoencoder (HR-VQVAE) that we have recently used with success for still image reconstruction [4]. Leveraging those two novel blocks, namely HR-VQVAE, and AST-PM, S-HR-VQVAE effectively tackles the video prediction task in three steps: In the first step, the input video frames are encoded to a continuous latent space and then mapped to discrete representations through HR-VQVAE, with each latent vector, in each layer in the model, assigned to a codeword in a codebook. The key property of this model is the strict hierarchy imposed between codebooks belonging to different layers, producing extremely compact and efficient discrete representations. In the second step, we predict future events in latent rather than image space. To perform this prediction, we use spatiotemporal modeling (the proposed AST-PM), where the distribution of the discrete latent representations for a particular location in the current frame is conditioned on the representations for neighboring locations both in space and time. In the third and final step, the predicted discrete representations are used by the HR-VQVAE decoder to generate the corresponding frame. Normally, HR-VQVAE and AST-PM may be trained independently. However, we also propose a novel joint training scheme to optimize HR-VQVAE and AST-PM together and show that this improves video prediction. We argue that the reason for the improved performance is that AST-PM and the decoder of HR-VQVAE are trained in such a way as to optimize both the predicted quantized latent representation for future frames as well as the reconstruction of future frames in image space.

Our contributions can be summarized as follows:

- S-HR-VQVAE, a novel technique for video prediction, is proposed. This includes a hierarchical vector quantized encoding scheme and a spatiotemporal autoregressive model of the latent representations. This model allows the capture of different levels of abstraction in a sequence of video frames, thus resulting in a compact but effective representation of the task.
- A novel loss function to jointly train the components of S-HR-VQVAE (HR-VQVAE and AST-PM) with further improvements of the prediction performance.
- State-of-the-art results on several challenging video prediction tasks, namely KTH Human Action [5], TrafficBJ [6], Human3.6M [7] and Kitti [8].

II. Related Work

A. Spatiotemporal Modeling

Hu et al. [9] introduced disentangled representation net (DrNet) for spatial feature modeling in single video frames, neglecting temporal information. Motion-content network (McNet) [10] and mutual suppression network (MsNet) [11] addressed motion and content separately, overlooking joint correlations. Convolutional long short-term memory (ConvLSTM) [12] aimed at capturing both spatial and temporal correlations but struggled with long-term dependencies and scalability. To overcome ConvLSTM's limitations, Wang et al. proposed predictive recurrent neural network (PredRNN) [13], which, despite improvements, still faced challenges in modeling complex long-term dependencies. PredRNN++ [14] and PredRNN-V2 [15] aimed to enhance PredRNN's performance by incorporating hierarchical recurrent structures. Eidetic 3D LSTM (E3D-LSTM) [16] was introduced to jointly model spatial and temporal dynamics. Su et al. [17] improved efficiency using low-rank tensor factorization, while robust spatiotemporal LSTM (R-ST-LSTM) [18] and memory in memory (MIM) [19] demonstrated performance improvements in long-term frame prediction tasks. The simple video prediction model (SimVP) [20] showed significant improvement over RNN-based models but struggled with encoding long-term dynamics, making accurate future prediction challenging. Chang et al. [21] introduce hierarchical semantic separation in video prediction using a spatiotemporal encoding-decoding scheme and residual predictive memory called STRPM. This scheme separates spatial and temporal information with independent encoders, preserving distinct features and improving high-resolution video predictions. The STRPM refines the separation by focusing on inter-frame residuals for more accurate future predictions. However, STRPM's reliance on residual inter-frame motion can oversimplify complex dynamics, and its implicit hierarchy may limit its ability to capture fine-grained spatiotemporal details compared to models with explicit multi-layered hierarchies.

To address the spatiotemporal challenge, S-HR-VQVAE leverages our proposed AST-PM module. In this module, causal convolutions in time and spatiotemporal self-attention are used to model the spatiotemporal correlations on the quantized codes level. Moreover, AST-PM operates on the latent discrete representations produced by the HR-VQVAE module instead of using pixels directly.

B. High Dimensionality

The above spatiotemporal methods rely on complex modeling, which hampers scalability, especially with the high dimensionality of video data. Hsieh et al. [22] addressed this by dividing frames into patches and predicting their evolution over time using a recurrent convolutional neural network (rCNN) [23]. Jun-Ting et al. [24] proposed the decompositional disentangled predictive autoencoder (DDPAE) framework, automatically breaking down high-dimensional videos into components with low-dimensional temporal dynamics. Xue et al. [25] proposed a variational autoencoder (VAE) [26] model to generate a distribution of next frame predictions. Oliu et al. [27] utilized a folded recurrent neural network (fRNN) with a gated recurrent unit (GRU) for bidirectional information flow, enabling state sharing between the encoder and decoder. Variational 3D ConvLSTM (V-3D-ConvLSTM) [28] combined variational encoder-decoder and 3D-ConvLSTM techniques. [29] developed thr video prediction Transformer (VPTR), an attention-based encoder-decoder, to learn local spatiotemporal representations while simplifying the model.

Compared to the aforementioned methods, S-HR-VQVAE can effectively manage the high dimensionality of video data, leveraging the hierarchical structure inside the vector quantization module, which efficiently compresses each video frame, as demonstrated in the experimental section.

C. Blurry Predictions

As reported in [30], video prediction solutions quite often rely on RNNs, VAEs, and their variants (e.g., variational RNNs - VRNNs [31]) resulting in blurry predictions. Two main strategies have emerged to address this issue: (i) Latent variable methods that explicitly model underlying stochasticity and (ii) Adversarially-trained models that aim to produce more natural images. In [32], the authors instead aimed to investigate stochastic models for video prediction using the VAE framework. Given the recent advances in generative adversarial networks (GANs), researchers have also explored alternative techniques, such as VAE-GANs [30], [33] for video frame prediction. VAE-GANs allow capturing stochastic posterior distributions of videos while making it feasible to model the spatiotemporal joint distribution of pixels. However, such methods often suffer from the problem of mode collapse and unrealistic predictions [33], [34].

S-HR-VQVAE combats image blurring thanks to the temporal model leveraging the hierarchical codebook representation.

This allows for an increase in the quantization granularity without resulting in blurry images. In fact, despite the lossy nature of the compressed encoding, our experiments clearly demonstrate that the original video can be reconstructed with a high degree of fidelity through the latent representations.

D. Physical Characteristics

To leverage physical characteristics, some methods focus on pixel-level representations. For example, De Brabandere et al. [35] introduced the dynamic filter network (DFN), which learns local spatial transformations from flow information. Finn et al. [36] proposed convolutional dynamic neural advection (CDNA), a model that predicts object motion and pixel motion distributions from previous frames. In another approach [37], a system was developed to predict optical flows between future and past frames. Berg et al. [38] utilized backward content transformation via a 6-parameter affine model to learn future-to-past relationships. Villegas et al. [10]] employed LSTM to independently model pixel-level images for spatial layout and temporal dynamics, simplifying prediction tasks. Guen et al. introduced the Physical dynamics network (PhyDNet) [39] to separate physical dynamics from other factors, yielding notable improvements. A motion-based modeling technique (MotionRNN) [40] decomposes motions into transient variations and trends, utilizing RNN-based models like ConvLSTM, PredRNN, and E3D-LSTM for prediction. Lee et al. [41] proposed a long-term motion context memory (LMC-memory) model for considering long-term motion context in future frame prediction. However, these methods primarily address physical characteristics, overlooking challenges like high dimensionality, blurry predictions, and spatiotemporal modeling in video prediction.

S-HR-VQVAE does not explicitly model physical characteristics. Nonetheless, the modularity of the hierarchical vector quantization block allows S-HR-VQVAE to implicitly model physical characteristics. In fact, latent representations are decomposed into a hierarchy of discrete codes, separating high-level global information (e.g., static background) from details (e.g., fine texture or small motions). Since the latent representations are decomposed into different layers of hierarchical residual codes, the proposed AST-PM can exploit spatiotemporal dependencies that are different for different levels of detail. For example, the background evolves slowly in time; whereas, the foreground object may move quickly. Similarly, within the foreground object, some details, such as hands and arms, may exhibit different movement patterns compared to the body. In sum, the combination of HR-VQVAE and AST-PM allows the modeling of physical characteristics, improving accuracy while reducing complexity.

III. Theoretical Background

Variational autoencoders (VAEs) and vector quantized VAEs (VQVAEs) have been used for many applications for their inherent representation capabilities [42], [43]. Focusing on image processing applications, our primary focus, an input image is represented as a tensor $\mathbf{x} \in \mathbb{R}^{H_I \times W_I \times D_I}$ of height H_I, width W_I and D_I color channels. VQVAE first maps the input image $\mathbf{x}$ to a continuous latent vector $\mathbf{z} \in \mathbb{R}^{H \times W \times D}$ through a non-linear encoder: $\mathbf{z} = E(\mathbf{x})$. Next, each element $\mathbf{z}_{hw} \in \mathbb{R}^D$, with $h \in [1, H]$, and $w \in [1, W]$, in the continuous latent vector $\mathbf{z}$ is quantized to the nearest codebook vector (i.e., a codeword) $\mathbf{e}_k \in \mathbb{R}^D,\ k \in 1, ..., m$ by

$$\text{Quantize}(\mathbf{z}_{hw}) := \mathbf{e}_k \text{ where } k = \arg\min_j \|\mathbf{z}_{hw} - \mathbf{e}_j\|_2. \quad (1)$$

The quantized vectors corresponding to each element $\mathbf{z}_{hw}$ are then recombined into the continuous representation $\mathbf{e} \in \mathbb{R}^{H \times W \times D}$ to form the input of the decoder that reconstructs the input image using a transformation $\mathcal{D}(\cdot)$. The loss function $\mathcal{L}(.)$ aims at minimizing the reconstruction error $\|\mathbf{x} - \mathcal{D}(\mathbf{e})\|_2$ whilst minimizing the quantization error $\|\mathbf{z} - \mathbf{e}\|_2$ as follows

$$\mathcal{L}(\mathbf{x}, \mathcal{D}(\mathbf{e})) = \|\mathbf{x} - \mathcal{D}(\mathbf{e})\|_2^2 + \|\text{sg}[\mathbf{z}] - \mathbf{e}\|_2^2 + \beta\|\text{sg}[\mathbf{e}] - \mathbf{z}\|_2^2, \quad (2)$$

where sg(.) is a stop-gradient operator cutting gradient flow during backpropagation, and β is a hyperparameter governing the stability of encoder output latent vectors.

In [44] a multi-layer version of VQVAE was proposed. However, the representations at different levels in the architecture were not related hierarchically.

IV. Proposed Method

In [4], we introduced a truly hierarchical version of VQVAE (HR-VQVAE) that is one of the building blocks of the video prediction method proposed in this work. HR-VQVAE deals with limitations in techniques such as VQVAE, e.g., codebook collapse and non-locality in codewords' indices. In HR-VQVAE, each layer captures residual information that is not properly modeled by the preceding layers, and the codebooks at different layers are constrained by a strict hierarchy.

Fig. 1 shows the proposed framework. Given T input frames $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ in a video, the goal is to predict the following S frames $(\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+S})$ in three steps. First, the input frames are encoded into a discrete latent representation using HR-VQVAE. Next, a novel autoregressive spatiotemporal predictive model (AST-PM) is proposed to predict new discrete latent variables of future frames based on the latent variables for previous frames. Finally, the HR-VQVAE decoder is used to generate the new frames from the latent variables obtained by AST-PM. The proposed approach is referred as sequential hierarchical residual learning vector quantized variational autoencoder (S-HR-VQVAE).

A. Step 1: Frame Encoding

In the first step, each frame $\mathbf{x} \in \mathbb{R}^{H_I \times W_I \times D_I}$ is encoded using HR-VQVAE into a discrete latent representation. HR-VQVAE first encodes the frame into a continuous vector $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^{H \times W \times D}$. These vectors are then iteratively quantized into n hierarchical layers of discrete latent embeddings. Assuming that the first layer has a single codebook of size M, the second layer has M independent codebooks of size M (for a total of M^2 codewords), and so on. A generic layer i has thereby M^{i-1} codebooks of size M, for a total of M^i codewords. However, only one of those codebooks is used in each layer depending on which codewords were chosen in the

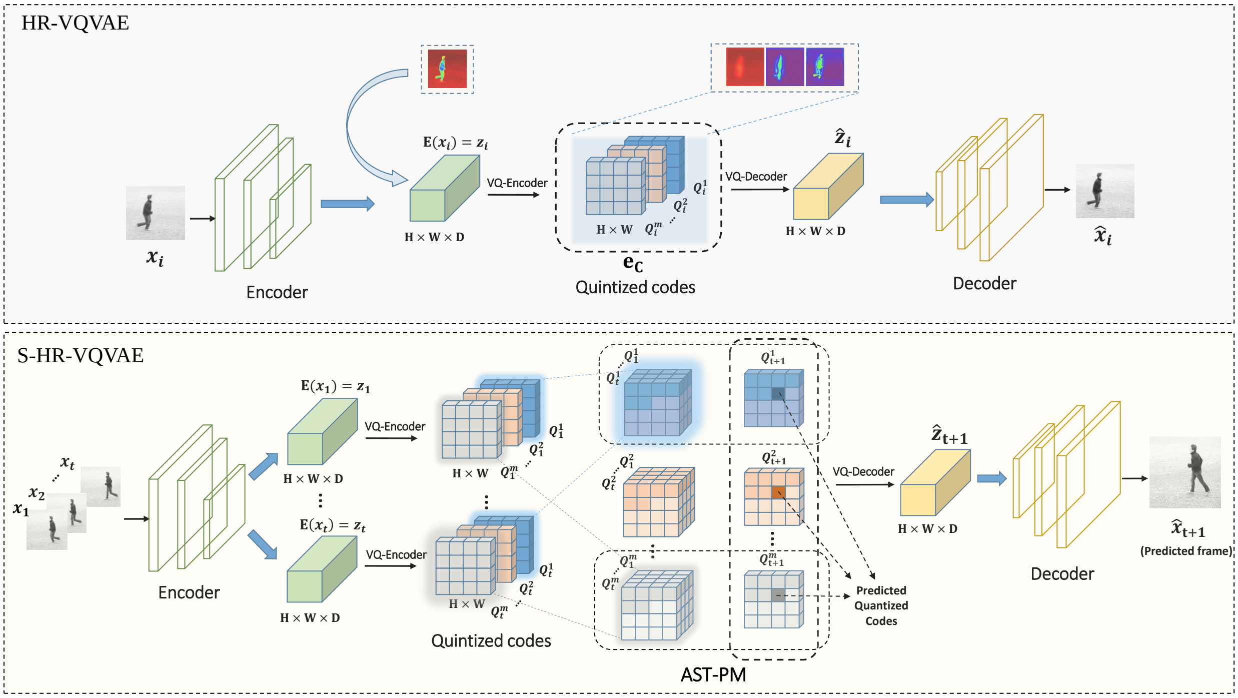


Fig. 1: Top: The HR-VQVAE module for hierarchical vector quantization, where each frame i is encoded into m hierarchical layers of quantized values $\left(Q_i^1, \ldots, Q_i^m\right)$. Bottom: Illustration of the S-HR-VQVAE for video prediction, which combines HR-VQVAE with the AST-PM model. AST-PM predicts the indices of quantized values in both spatial and temporal dimensions, where each index at time $t+1$ is predicted by accessing only its preceding indices—those located above and to the left in a raster-scan spatial order and those before $t+1$ in the temporal domain.

previous layers. In each layer i, the codebook is optimized to minimize the error between codewords $\mathbf{e}_k^i \in \mathbb{R}^D$ and elements $\boldsymbol{\xi}_{hw}^{i-1} \in \mathbb{R}^D$ of the residual error from the previous layer[1]

$$\text{Quantize}^i(\boldsymbol{\xi}_{hw}^{i-1}) := \mathbf{e}_k^i \text{ where } k = \arg\min_j \|\boldsymbol{\xi}_{hw}^{i-1} - \mathbf{e}_j^i\|_2, \quad (3)$$

and $\mathbf{e}_k^i$ belongs to one of the possible codebooks $C_i(t)$ for layer i. Which codebook is used is determined by the codeword $\mathbf{e}_t^{i-1}$ selected at the previous layer. Within each layer, the codewords $\mathbf{e}_k^i$, for each element $\boldsymbol{\xi}_{hw}^{i-1}$ of the residual, are combined to form the tensor $\mathbf{e}^i \in \mathbb{R}^{H\times W\times D}$. Across the different layers, the tensors $\mathbf{e}^i$ are then summed to form the "combined" discrete representation $\mathbf{e}_C$. When HR-VQVAE is used to reconstruct single images, $\mathbf{e}_C$ is fed into the decoder to reconstruct the image as $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{e}_C)$, and the corresponding objective function is used to train the system

$$\mathcal{L}(\mathbf{x}, \mathcal{D}(\mathbf{e}_C)) = \|\mathbf{x} - \mathcal{D}(\mathbf{e}_C)\|_2^2 + \|\text{sg}[\boldsymbol{\xi}^0] - \mathbf{e}_C\|_2^2 + \beta_0\|\text{sg}[\mathbf{e}_C] - \boldsymbol{\xi}^0\|_2^2 + \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\xi}^{i-1}, \mathbf{e}^i), \quad (4)$$

with

$$\mathcal{L}(\boldsymbol{\xi}^{i-1}, \mathbf{e}^i) = \|\text{sg}[\boldsymbol{\xi}^{i-1}] - \mathbf{e}^i\|_2^2 + \beta_i\|\text{sg}[\mathbf{e}^i] - \boldsymbol{\xi}^{i-1}\|_2^2. \quad (5)$$

The β_i are hyperparameters that control the reluctance to change the code corresponding to the encoder output. The main goal of Eqs. 4 and 5 is to make a hierarchical mapping of input data in which each layer of quantization extracts residual information from its bottom layers.

In the proposed S-HR-VQVAE, we do not reconstruct images directly. The indices to the codewords $\mathbf{e}^i$ are, instead, used as latent representations for each input frame in the video and each layer in the system and are input to the video prediction steps described below. We call these indices for layer i, $Q^i \in [1, M]^{H\times W}$, with M the size of each codebook. The difference is clarified in Figure 1, where the image reconstruction case is represented in the top panel, and the video prediction is depicted in the bottom panel.

B. Step 2: Spatiotemporal Latent Representation Prediction

In the second step, the indices $(Q_1^i, \ldots, Q_T^i)$ of the codewords $(\mathbf{e}_1^i, \ldots, \mathbf{e}_T^i)$, obtained from each layer i of HR-VQVAE from the input frames $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$, are used to predict the indices $(Q_{T+1}^i, \ldots, Q_{T+S}^i)$ of the codewords $(\mathbf{e}_{T+1}^i, \ldots, \mathbf{e}_{T+S}^i)$ for S future frames, with the goal of predicting the S next future frames $(\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+S})$.

To this end, we propose a probabilistic autoregressive spatiotemporal predictive model (AST-PM). AST-PM takes discrete indices of the latent representations as input and predicts future indices. Because HR-VQVAE is hierarchical, it greatly simplifies predicting spatial and temporal information, which allows our model to focus on the most important parts of the frames in both space and time. Accordingly, the model predicts future codeword indices $\hat{Q}t > T$ using

[1] For the first layer, $\boldsymbol{\xi}_{hw}^0 \equiv \mathbf{z}_{hw}$.

the codeword indices from previous times $(Q_1^i, \ldots, Q_T^i)$. To explain our probabilistic model, we first arrange the elements of $Q_t^i \in [1, M]^{H \times W}$ from left to right and top to bottom using a linear index $v_k \in [1, HW]$. We use the notation $vj < k$ to refer to any element of Q_t^i that is to the left or above v_k. Given this notation, the probabilistic model can be written as:

$$p(\hat{Q}_{t+1}^i(v_k)) = \prod_{j=1}^{H \times W} p(\hat{Q}_{t+1}^i(v_{j<k}) | Q_1^i(v_{j<k}), \ldots, Q_t^i(v_{j<k})), \tag{6}$$

where $\hat{Q}_t^i$ represents the predicted quantized discrete codes of layer i obtained from the t^{th} frame. This behavior is achieved by using convolutional masks to limit the information used during prediction. The convolutional masks restrict the convolutions to retrieve only spatial information from the left and above each index. For the temporal dimension, convolutions are limited to previous time steps by masking out present and future timesteps. This strategy is implemented using multi-head attention layers similar to those in [44]. However, in our case, the attention is applied to 3D voxels. The loss function of AST-PM is as follows

$$\mathcal{L}_p(p(Q_{t>T}^i), p(\hat{Q}_{t>T}^i)) = \\ -\frac{1}{H \times W} \sum_{j=1}^{H \times W} \sum_{m=1}^{M} p(Q_{t>T}^i)[j, m] * \log p(\hat{Q}_{t>T}^i)[j, m], \tag{7}$$

C. Step 3: Frame Generation

Once the quantization indices $\hat{Q}_t^i$ for each layer i and each time step $t \in [T + 1, T + S]$ have been estimated by the AST-PM, the corresponding quantized representation $\hat{\mathbf{z}}_t \in \mathbb{R}^{H \times W \times D}$ can be computed by codebook access $\mathbf{e}_{C(t)} = \sum_{i=1}^{m} e_t^i$ (see Section IV-A).

Finally, the predicted quantized codes are decoded to sequences of frames using the HR-VQVAE decoder $\mathcal{D}(.)$

$$(\hat{\mathbf{x}}_{T+1}, ..., \hat{\mathbf{x}}_{T+S}) = (\mathcal{D}(\hat{\mathbf{z}}_{T+1}), ..., \mathcal{D}(\hat{\mathbf{z}}_{T+S})), \tag{8}$$

where the and $\hat{\mathbf{z}}_{t>T}$ and $\hat{\mathbf{x}}_{t>T}$ represent the predicted latent representations and frames, respectively.

D. Disjoint and Joint Training

HR-VQVAE and AST-PM in the combined model described above can be trained independently. In this case, we first train HR-VQVAE according to Eq. 4 to predict each frame $\mathbf{x}_i$ in the video independently of the others. We obtain a sequence of latent representations $(Q_1^i, \ldots, Q_T^i, Q_{T+1}^i, \ldots, Q_{T+S}^i)$ for each layer in HR-VQVAE and for the complete sequence of frames. The AST-PM can the been trained to predict the sequence $(Q_{T+1}^i, \ldots, Q_{T+S}^i)$ given the input sequence $(Q_1^i, \ldots, Q_T^i)$, by optimizing Eq. 7. In the test phase, we use the predictions of AST-PM in combination with the HR-VQVAE decoder to predict unseen video frames, making sure that the combined model only has access to $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ when predicting $(\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+S})$.

Following this training procedure, the decoder in HR-VQVAE is exclusively optimized to deal with the uncertainty introduced by the encoder of HR-VQVAE, which means that the decoder is optimized solely for reconstructing the original input frame. However, when we reconstruct the frames $(\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+S})$, we also need to deal with the uncertainty introduced by the AST-PM predictions. In fact, the AST-PM uncertainty refers to the mismatch between the predicted latent spaces of future frames and the actual latent space of future frames. This indicates that the HR-VQVAE decoder block and the AST-PM block operate independently, without being aware of the uncertainties introduced by the other block. In an attempt to address this issue, we propose to optimize the AST-PM and the HR-VQVAE decoder jointly. Therefore, we proposed a joint training in Eq. 9 which includes two distinct objectives: the loss for the HR-VQVAE decoder, represented by the first term of Eq. 4, and the AST-PM loss in Eq. 7. The corresponding multi-objective loss is:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_p + \lambda \|\mathbf{x}_t - \mathcal{D}(\mathbf{e}_{C(t)})\|_2^2, \tag{9}$$

where λ is a hyperparameter that controls the effect of the reconstruction loss on the joint training. In this case, during training, HR-VQVAE only produces the latent representations $(Q_1^i, \ldots, Q_T^i)$ for the input frames $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$. The latent representations $(Q_{T+1}^i, \ldots, Q_{T+S}^i)$ for the frames $(\mathbf{x}_{T+1}, \ldots, \mathbf{x}_{T+S})$ are predicted by AST-PM and then used to train the HR-VQVAE decoder.

V. Experiments

A. Datasets

We conducted experiments using four different challenging datasets. Table I presents a summary of corresponding statistics, including the number of training samples (#Train), the number of test samples (#Test), image resolution represented as (H, W, C), input sequence length indicated as T, and predicted sequence length referred to as $\hat{T}$.

The **KTH Human Action** dataset [5] is a moving image dataset with a resolution of 160×120 pixels that contains six types of human actions, including walking, jogging, running, boxing, hand waving, and hand clapping. The dataset comprises 25 human subjects performing actions in four different scenarios. For our experiments, we followed [15], resized the video frames down to 128×128, and split the dataset into two subsets: (i) a training set, consisting of the first 16 subjects, and (ii) a test set, containing the remaining subjects.

The **TrafficBJ** is a collection of taxicab GPS data and meteorological data recorded in Beijing [6]. Each frame in TrafficBJ has 32×32 pixels, including the traffic flow entering and leaving the same district. We normalized the data to $[0, 1]$ and follow the experimental settings as [45].

The **Human3.6M** dataset [7] consists of 3.6 million samples capturing diverse human activities. Similar to previous papers [19], [20], [39], we focus on the "walking" scenario.

The **Kitti** dataset [8] was created through real traffic scenario collections by specially equipped vehicles, a joint effort by Germany's Karlsruhe Institute of Technology and the Toyota Institute of Technology in the United States. We employ Kitti using three scenarios: road, city, and residential, resulting in 57 videos for a training set and 4 for a test set.

TABLE I: Dataset statistics. #Train and #Test indicate the number of samples for the training and test set, respectively. Each input sequence consists of T, and the output sequence consists of $\hat{T}$ frames with shape (H, W, C).

Dataset	**#Train**	**#Test**	**(H, W, C)**	T	$\hat{T}$
TrafficBJ [6]	19,627	1,334	(32, 32, 2)	4	4
KTH [5]	5,200	3,167	(128,128, 3)	10	20
Human3.6M [7]	2,624	1,135	(128, 128, 3)	4	4
Kitti [8]	40,783	1,963	(128, 128, 3)	4, 5	5

TABLE II: Configuration details for S-HR-VQVAE.

	KTH & Human3.6M & Kitti			**TrafficBJ**		
Input size	128×128			32×32		
Bit rate	8			8		
Latent size	$32 \times 32 \times 8$			$16 \times 16 \times 4$		
Quantized size	32×32			16×16		
#Layers	1	3	9	1	3	6
Codebook size	512	8	2	64	4	2
#Codewords	512	{8, 64, 512}	{2, 4,..., 512}	64	{4, 16, 64}	{2, 4,..., 64}

Finally, it is important to note that 5% of the training set was reserved as a validation set, which was used specifically for fine-tuning the hyperparameters.

B. Experimental Setup

Table II lists some details about S-HR-VQVAE architecture for tackling the datasets. *Input size* refers to the initial resolution of the video frames. *Latent size* corresponds to the continuous latent representation in HR-VQVAE. *Quantized latent size* to the quantized representation in the model. We also provide additional information for the bit rate, number of hierarchy layers, codebook size, and number of codewords.

The proposed S-HR-VQVAE was trained on sequences consisting of 10 consecutive frames to predict 20 future frames for KTH Human Action, and also trained on 4 consecutive frames to predict 4 future frames for both TrafficBJ and Human3.6M datasets, which is a common practice for the tasks. In addition, for the Kitti dataset, we focused on two specific settings: (i) 4 input frames and 5 predicted frames and (ii) 5 input frames and 5 predicted frames. In all experiments, the model is trained using the Adam optimizer [46], and the learning rate is set to 0.0003 for both HR-VQVAE encoder-decoder and AST-PM. Besides, λ in Eq. 9 is set to 0.11.

C. Metrics

We report results adopting metrics that are commonly used in the literature, namely: peak signal-to-noise ratio (PSNR) [47], structural similarity index measure (SSIM) [48], learned perceptual image patch similarity (LPIPS) [49], frechet video distance (FVD) [50], mean square error (MSE), and mean absolute error (MAE). PSNR, SSIM, LPIPS, MSE, and MAE are all image quality metrics but differ in their characteristics. PSNR focuses on signal-to-noise ratio, SSIM considers structural similarity, MSE and MAE measure pixel-wise differences, and LPIPS aims to capture perceptual similarity based on deep neural networks. FVD, on the other hand, is a comprehensive video quality metric employed to assess the quality of generated videos. This evaluation is achieved by quantifying the feature distribution gap between real and generated videos, which effectively captures both temporal inconsistencies and motion-related artifacts. Furthermore, FVD evaluates both the temporal coherence of video content and the quality of individual frames, offering a holistic perspective on video realism and overall coherence. All those metrics, however, have limitations. For example, PSNR, MSE, and MAE have been shown to have poor correlation with human perception [51], [52] and may not take into account higher-level semantic information, such as in action modeling. SSIM and LPIPS are more effective in capturing perceptual differences, but they may not be sensitive to all types of visual information: they may not be as effective at capturing differences in color or texture as at capturing differences in luminance and contrast [53]. fFVD tends to prioritize a video's spatial elements and may overlook the natural flow of its temporal dynamics [54]. Therefore, several metrics must be considered to better capture different aspects of the video prediction task and obtain a more comprehensive assessment of the methods' performance.

We report results according to all those metrics and include all available results for the related methods. Because of the limitations of these metrics, we also provide a qualitative assessment to verify whether the metrics have missed some important aspects of the video prediction task.

VI. Results

Results of the quantitative evaluation of the proposed method followed by a qualitative assessment are now presented. To better appreciate the effectiveness of the proposed technique, we have performed a systematic review of reported quantitative results of recent, state-of-the-art solutions.

The qualitative analysis is performed by observing the behavior of the proposed method on several video sequences, which is a common practice in the research field. However, while reviewing the literature, we noticed that different methods use different video sequences to visually demonstrate the quality of their approaches; furthermore, the source code is not available for all methods in the literature, which implies that different systems can not be compared on the same set of predefined video sequences. To overcome that issue, we first selected video sequences common among different techniques in the literature. Then, we evaluated our S-HR-VQVAE on those selected examples and grouped the results accordingly. To the best of our knowledge, this is the first time that such a systematic comparison has been carried out.

We also provide results for other aspects of the proposed S-HR-VQVAE, including reconstruction capability in (i) blur mitigation, (ii) noise removal, and (iii) compression.

A. Quantitative Analysis

In this study, we assess the performance of state-of-the-art video prediction methods on different datasets, providing a comprehensive overview of the advancements in the field. In particular, Tables III, IV and V list state-of-the-art methods from 2015 to 2023 in a chronologically ascending order, highlighting thereby the evolution of the techniques over the years.

TABLE III: Results on KTH Human Action dataset. S-HR-VQVAE with 3 layers was used with disjoint and joint training.

	KTH Human Action (10 → 20)				
Method	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**#Params**	**FLOPs**
ConvLSTM (2015) [12]	23.01	0.704	0.156	16.60M	1,468G
DFN (2016) [35]	27.26	0.794	×	×	×
CDNA (2016) [36]	23.75	0.752	×	×	×
DrNet(2017) [9]	25.56	0.764	×	23.30M	×
PredRNN (2017) [13]	27.55	0.839	0.167	23.85M	2,800G
McNet (2018) [10]	25.95	0.804	×	3.50M	×
MsNet (2018) [11]	27.08	0.876	×	3.20M	×
fRNN (2018) [27]	26.12	0.771	×	×	×
PredRNN++ (2018) [14]	28.62	0.888	0.229	15.40M	4,162G
E3D-LSTM (2019) [16]	27.92	0.893	×	41.94M	214.0G
MIM (2019) [19]	27.78	0.902	0.188	37.37M	1,099G
Conv-TT-LSTM (2020) [17]	28.36	0.907	0.133	39.8M	×
PhyDNet (2020) [39]	28.69	×	0.188	3.10M	**93.6G**
Jin et al. (2020) [55]	29.85	0.893	0.118	×	×
LMC-Memory (2021) [41]	28.61	0.894	0.133	×	×
V-3D-ConvLSTM (2021) [28]	28.31	0.866	×	12.90M	×
R-ST-ConvLSTM (2022) [18]	28.99	0.854	×	×	×
SimVP (2022) [20]	**33.72**	0.905	×	22.30M	125.6G
PredRNN-V2 (2023) [15]	28.37	0.838	0.139	23.86M	2,815G
VPTR (2023) [29]	26.96	0.879	0.076	162.48M	×
NPVP (2023) [56]*	27.66	0.909	**0.066**	×	×
S-HR-VQVAE-disjoint (ours)	28.43	0.863	0.130	**1.14M**	94.1G
S-HR-VQVAE-joint (ours)	28.49	**0.910**	0.093	**1.14M**	95.8G

* NPVP resized KTH samples to 64 × 64 instead of standard 128 × 128.
(↑) means higher is better and (↓) means lower is better.

On the KTH Human Action task, PSNR and SSIM are reported by all competing techniques; whereas, LPIPS is provided for only a few methods. On TrafficBJ and Human3.6M tasks, we report MSE, MAE, and SSIM as in [19], [20], [39]. Finally, on the Kitti dataset, we report SSIM, LPIPS, FVD, and PSNR. Here we report our results in order of complexity of the task (from KTH Human action to Kitti).

For the **KTH Human Action** task, from Table III, it is evident that the proposed S-HR-VQVAE outperforms all methods, up to fRNN, in *all* reported metrics. Among methods introduced after fRNN, S-HR-VQVAE outperforms PredRNN++ on two metrics out of three, E3D-LSTM on all, Conv-TT-LSTM on all, PhyDNet on one out of two, Jin et al. [55] on two out of three, LMC-Memory on two out of three, V-3D-ConvLSTM across all, R-ST-ConvLSTM on one out of two, SimVP on one out of two, PredRNN-V2 on all, VPTR on two out of three, and NPVP on two out of three. It can also be seen from Table III that SimVP has the overall best PSNR, but our method outperforms it and achieves the best result in terms of SSIM. For LPIPS, NPVP has the best performance; however, the method is outperformed by our method both in terms of PSNR and SSIM, despite NPVP downsampling video frames to 64×64 rather than the typical 128×128. It is noteworthy that S-HR-VQVAE achieves those results with a significantly lower number of parameters with respect to all other methods and ranks second in terms of computational efficiency (FLOPs).

On the **TrafficBJ** task, as detailed in Table IV, S-HR-VQVAE exhibits exceptional performance, outperforming existing state-of-the-art methods on all evaluation metrics. Our model particularly stands out by significantly outperforming methods such as PhyDNet and SimVP, achieving the highest scores across all metrics. A similar trend is observed on the challenging task of **Human3.6M**, where S-HR-VQVAE again outperforms the current state-of-the-art approaches, leading in all evaluation metrics, especially in FVD (better temporal modeling) and FLOPs (computational efficiency).

The performance of the S-HR-VQVAE on the challenging task of the **Kitti** dataset is detailed in Table V. Unlike other tasks such as KTH Human Action, TrafficBJ, and Human3.6M, where the background is static, the Kitti dataset introduces a unique challenge with its dynamic and complex environments. This complexity comes from the challenging driving scenes, where both the foreground and background are in motion. This requires the prediction model to accurately handle multiple moving elements and rapidly changing landscapes, a significant shift from tasks where movement is mainly due to a single object against a constant background.

For the Kitti (4 → 5) task, S-HR-VQVAE has demonstrated remarkable improvement over traditional models like PredRNN and McNet and more recent approaches such as NPVP. It not only obtains better performance in SSIM, showing the best perceptual quality of predictions but also achieves the lowest LPIPS and a significantly better FVD, indicating superior performance in capturing both spatial and temporal aspects of the scenes. Similarly, for the Kitti (5 → 5) task, S-HR-VQVAE significantly outperforms other state-of-the-art models such as PhyDNet, LMC-Memory, MotionRNN, and MIMO. It achieves higher SSIM and PSNR values, which indicates that it not only captures higher structural similarities between the predicted and actual frames but also maintains high-quality predictions across various frames. The lower LPIPS also shows further evidence of S-HR-VQVAE's capability to preserve more accurate textural and detail-oriented features that are critical in dynamic scenes.

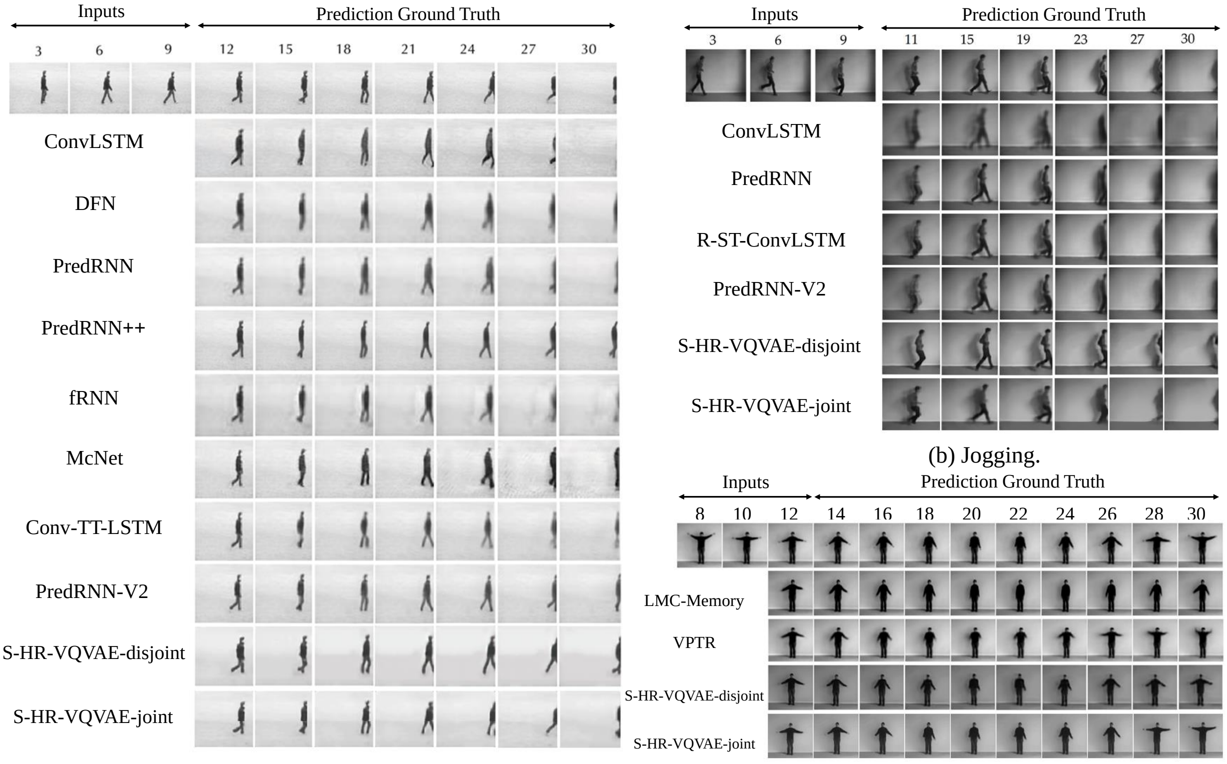
Referring to Tables III, IV and V, we can observe that the different metrics improve over the years. Also, it can be argued that starting from 2018, all methods are quite competitive with one another, and it is not possible to indicate a single technique that performs the best on the video prediction task across all metrics. Indeed, when we attempt to determine the best method, it can be seen that methods performing best in one metric are usually outperformed by other methods in other metrics, and therefore, it is essential to evaluate the results using all available metrics. From this analysis, we can conclude that although some of the state-of-the-art methods outperform our method on a single metric, S-HR-VQVAE is more robust across all metrics for the considered tasks, especially for the challenging tasks of Human3.6M and Kitti, where S-HR-VQVAE outperforms the state-of-the-art methods across all metrics. Ultimately, Tables III, IV, and V show the positive impact of jointly training HR-VQVAE and AST-PM across all datasets. While joint training introduces a slight increase in FLOPs compared to disjoint training, this marginal rise is outweighed by the significant benefits of joint training, particularly in improving spatiotemporal modeling, which contributes to the overall performance of the model.

The effectiveness of our approach can be further appreciated by considering the following qualitative analysis since objective metrics might not capture all aspects of the actual quality of the predicted sequences.

TABLE IV: Results on TrafficBJ and Human 3.6M. S-HR-VQVAE with 3 layers was used with disjoint and joint training.

	TrafficBJ (4 → 4)				Human3.6M (4 → 4)				
Method	**MSE×100↓**	**MAE↓**	**SSIM↑**	**FLOPs**	**MSE/10↓**	**MAE/100↓**	**SSIM↑**	**FVD↓**	**FLOPs**
ConvLSTM (2015) [12]	48.5	17.7	0.978	20.74G	50.4	18.9	0.776	28.4	347.0G
PredRNN (2017) [13]	46.4	17.1	0.971	42.40G	48.4	18.9	0.781	24.7	704.0G
PredRNN++ (2018) [14]	44.8	16.9	0.977	62.95G	×	×	×	×	1,033G
E3D-LSTM (2019) [16]	43.2	16.9	0.979	98.19G	46.4	16.6	0.869	23.7	542.0G
MIM (2019) [19]	42.9	16.6	0.971	64.10G	42.9	17.8	0.790	21.8	1,051G
PhyDNet (2020) [39]	41.9	16.2	0.982	5.60G	36.9	16.2	0.901	18.3	19.1G
MotionRNN (2021) [40]	×	×	×	×	34.2	14.8	0.846	18.3	49.5G
SimVP (2022) [20]	41.4	16.2	0.982	3.61G	31.6	15.1	0.904	×	197.0G
PredRNN-V2 (2023) [15]	45.6	16.8	0.980	42.63G	36.3	17.7	0.863	×	708.0G
S-HR-VQVAE-disjoint (ours)	41.5	16.2	0.985	**3.11G**	30.9	14.4	0.916	16.5	**16.7G**
S-HR-VQVAE-joint (ours)	**40.3**	**15.2**	**0.993**	4.07G	**30.4**	**12.4**	**0.939**	**15.2**	17.4G

(↑) means higher is better and (↓) means lower is better.



(a) Walking.

(b) Jogging.

(c) Handwaving.

Fig. 2: Comparison of S-HR-VQVAE with state-of-the-art methods on KTH Human Moving Action dataset over three sequences (a, b, and c) that are commonly reported in the literature. It should be noted that 10 frames (1-10 in the figures) are given as input, and the next 20 frames (11-30 in the figures) are predicted.

B. *Qualitative Analysis*

Figure 2 shows the predictions for different state-of-the-art methods and S-HR-VQVAE on the KTH Human Action dataset for three different activities: walking (panel a), jogging (panel b), and handwaving (panel c). In the hand wave activity, for example, hand movements are relatively fast, but S-HR-VQVAE can better predict the ground truth whilst avoiding blurry outputs, as shown in frames 28 and 30. In the walking task, most methods do not predict well the position of the body and the legs, except for our method, PredRNN, PredRNN++, and PredRNN-V2 (see frames 27 and 30, for example). However, our method produces sharper images and correctly predicts the location of both legs for these frames. Finally, for the jogging task, an overall better estimation of the location of the jogger is observed along with sharper images.

Figure 3 presents a qualitative analysis of the results obtained for TrafficBJ samples. To enhance the clarity of our comparisons, we include visualizations of the differences between the predictions and the corresponding ground truth images. S-HR-VQVAE demonstrates impressive performance in generating predicted frames when compared to the other models, as evidenced by the minimal intensity of differences observed. It is noteworthy that S-HR-VQVAE obtains the best result on all metrics for this task.

The qualitative analysis presented in Figure 4 reveals that S-HR-VQVAE generates more precise predictions for motion positions and object sizes. This observation underscores the efficacy of S-HR-VQVAE when applied to intricate real-world

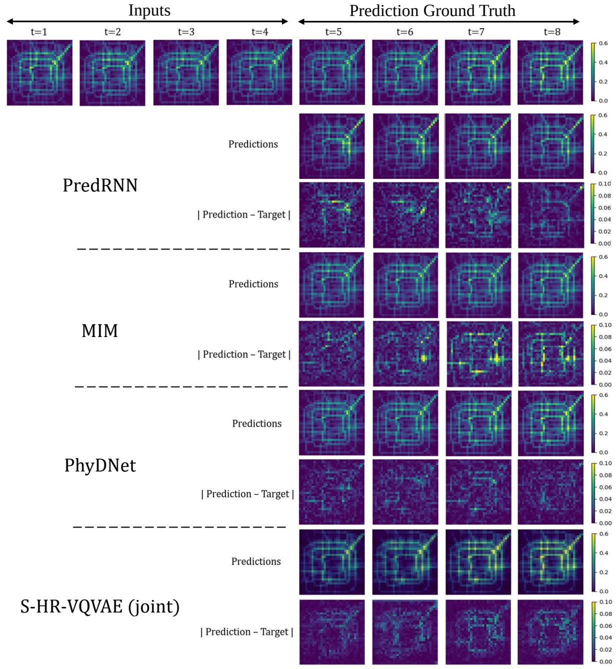


Fig. 3: Comparison of S-HR-VQVAE with state-of-the-art-methods on TrafficBJ dataset. It should be noted that 4 frames (1-4 in the figure) are given as input, and the next 4 frames (5-8 in the figure) are predicted.

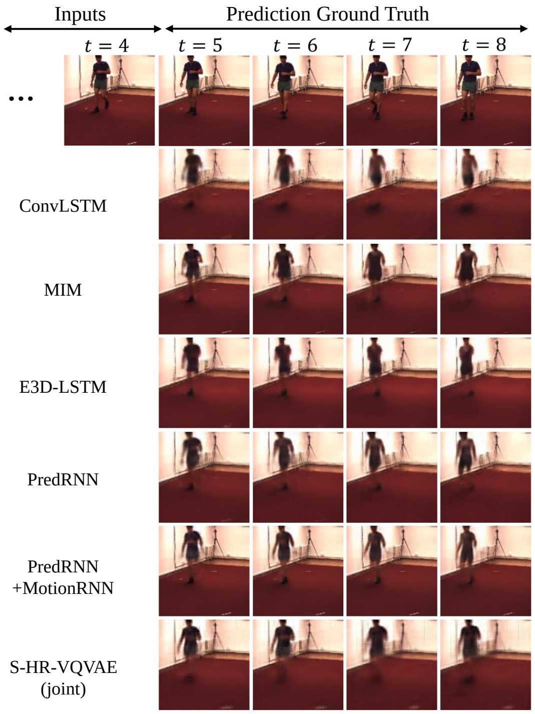


Fig. 4: Comparison of S-HR-VQVAE with state-of-the-art-methods on Human3.6M dataset. 4 frames (1-4 in the figure) are given as input, and the next 4 frames (5-8 in the figure) are predicted.

TABLE V: Results on Kitti dataset. S-HR-VQVAE with 3 layers was used with disjoint and joint training.

	Kitti (4 → 5)		
Method	**SSIM↑**	**LPIPS↓**	**FVD↓**
PredRNN (2017) [13]	0.475	0.629	×
McNet (2018) [10]	0.554	0.373	×
NPVP (2023) [56]	0.661	0.279	134.69
S-HR-VQVAE-disjoint (ours)	0.673	0.188	127.04
S-HR-VQVAE-joint (ours)	**0.692**	**0.164**	**121.84**
	Kitti (5 → 5)		
	SSIM↑	**LPIPS↓**	**PSNR↑**
PhyDNet (2020) [39]	0.674	0.403	19.159
LMC-Memory (2021) [41]	0.660	0.410	18.692
MotionRNN (2021) [40]	0.652	0.384	18.931
MIMO (2023) [57]	0.703	0.308	19.616
S-HR-VQVAE-disjoint (ours)	0.845	0.187	19.774
S-HR-VQVAE-joint (ours)	**0.861**	**0.114**	**21.877**

(↑) means higher is better and (↓) means lower is better.

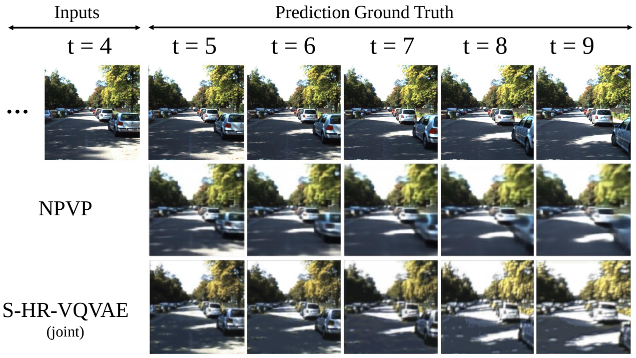


Fig. 5: Comparison of S-HR-VQVAE with the state-of-the-art method on the Kitti dataset, where 4 frames are given as input, and the next 5 frames are predicted.

datasets. S-HR-VQVAE better performance in predicting object positions and sizes can be attributed to the collaborative interaction between our spatiotemporal predictive model and the decoder, as stated in the objective function in Eq. 9.

The qualitative assessment on the Kitti dataset is depicted in Figure 5. From the visual analysis, it is evident that S-HR-VQVAE exhibits finer details, such as intricate shadow patterns, leaf textures on trees, and more precise car features, compared to NPVP. Moreover, S-HR-VQVAE significantly reduced blurry predictions compared to NPVP. These observations align well with the quantitative findings presented in Table V, where S-HR-VQVAE outperforms NPVP across all evaluation metrics: SSIM, LPIPS, and FVD. The higher SSIM score of S-HR-VQVAE indicates better structural similarity between predicted and ground truth frames, while the lower LPIPS value suggests reduced perceptual differences, underscoring the model's ability to generate more visually faithful predictions. Furthermore, the significantly lower FVD score of S-HR-VQVAE compared to NPVP highlights its superiority in capturing temporal consistencies and minimizing artifacts.

We can summarise the outcome of the qualitative analysis as follows: Although quantitative analysis is useful for understanding whether a sequence prediction technique is viable or not, objective measures by themselves may not

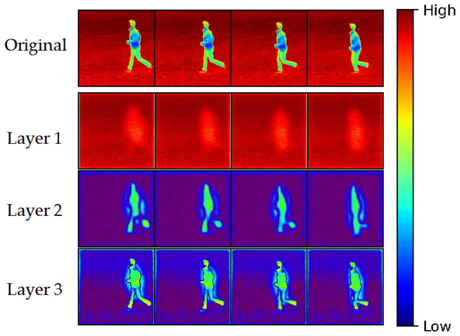


Fig. 6: Heatmap of reconstructions obtained from different layers of a 3-layer HR-VQVAE.

reveal the actual capability of a technique. State-of-the-art methods exhibit a varying sequence prediction quality across tasks, as observed, for example, in Figure 2 despite the good numerical results reported in Table III, whereas S-HR-VQVAE performs consistently across tasks. Moreover, the figure suggests joint training within our methodology leads to a significant enhancement in location prediction. This improvement is evident across the majority of the frames. Nevertheless, it is important to acknowledge that while this improvement in location prediction is evident, it appears to be accompanied by a minor reduction in image sharpness in the reconstructed frames. This observation may provide insights into the relatively modest quantitative improvements observed in our results following the incorporation of joint training.

VII. Discussion

A. Model Interpretability

To facilitate the interpretation of latent representations produced by the model, we present heatmaps over various layers of HR-VQVAE in Fig. 6. Each heatmap highlights regions of significance within the reconstructed latent representation. General information, i.e., background, is mainly captured in the first layer; the second layer focuses on the position of the foreground object, whereas the third layer is concerned with details of the moving objects.

B. Blur, Noise, and Compression

To gain more insights into the effectiveness of S-HR-VQVAE against blurriness, we artificially corrupt some video sequences by injecting Gaussian Blur (Fig. 7-a) and Fragment Blur (Fig. 7-b). The prediction results reported in those figures demonstrate that HR-VQVAE can successfully reduce blurriness while being able to reconstruct details in the images that were lost due to the blur effect. In addition to blur mitigation, HR-VQVAE is also robust to noise, as shown in Fig. 7-c, where accurate sequence prediction is attained although the input frames were artificially corrupted with additive noise at different SNR levels. Finally, we show the reconstruction of compressed images with two levels of compression ratio in Fig. 7-d, showcasing the HR-VQVAE's robustness against compression. HR-VQVAE robustness against blur, noise, and compression in sequence prediction is especially valuable in applications where the quality of the predicted video frames is critical, such as autonomous driving.

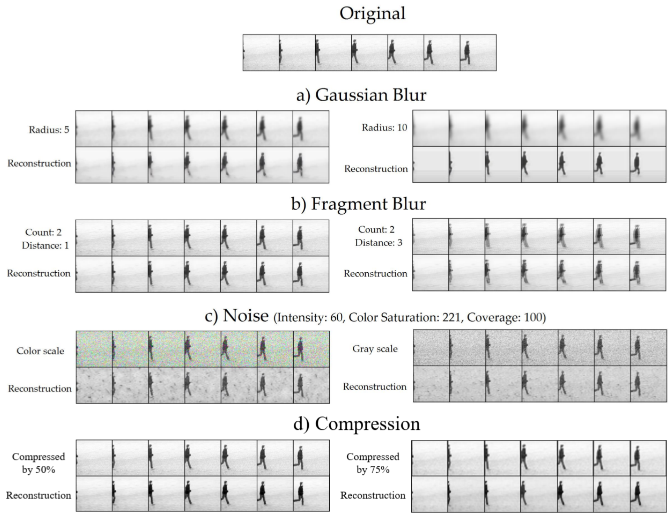


Fig. 7: Reconstructions by 3-layer HR-VQVAE. a) Gaussian Blur b) Fragment Blur c) Noise d) Compression. Zoom in to see more details.

VIII. Conclusion

In this study, we proposed a video prediction framework that combines the hierarchical vector quantization codebooks of the previously proposed HR-VQVAE with the novel autoregressive spatiotemporal predictive model (AST-PM). We call this method sequential HR-VQVAE (S-HR-VQVAE). We show how the proposed S-HR-VQVAE takes advantage of hierarchical frame modeling to model different levels of abstraction, enabling the system to capture both context and movements (details) in video frames with a fraction of the parameters used by competing models. We show by extensive experimental evidence on the KTH Human Action, TrafficBJ, Human3.6M, and Kitti tasks that the model is very competitive with the state-of-the-art in video prediction, outperforming the best methods, at least in a subset of the available metrics (PSNR, SSIM, LPIPS, FVD, MSE, and MAE) with significantly lower number of parameters. We provide a detailed analysis of the properties of the model, including an analysis of its internal representations and its behavior concerning blurry and noisy input frames. The proposed method is competitive for the video prediction task, in terms of performance, low complexity, and interpretability.

References

[1] V. Vukotić, S.-L. Pintea, C. Raymond, G. Gravier, and J. C. Van Gemert, "One-step time-dependent future video frame prediction with a convolutional encoder-decoder neural network," in *International conference on image analysis and processing*. Springer, 2017, pp. 140–151.

[2] C. Lu, M. Hirsch, and B. Schölkopf, "Flexible spatio-temporal networks for video prediction," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[4] M. Adiban, K. Stefanov, S. M. Siniscalchi, and G. Salvi, "Hierarchical residual learning based vector quantized variational autoencoder for image reconstruction and generation," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.

[5] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.

[6] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, and T. Li, "Predicting citywide crowd flows using deep spatio-temporal residual networks," *Artificial Intelligence*, vol. 259, pp. 147–166, 2018.

[7] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4417–4426.

[10] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *5th International Conference on Learning Representations, ICLR 2017*. International Conference on Learning Representations, ICLR, 2017.

[11] J. Lee, J. Lee, S. Lee, and S. Yoon, "Mutual suppression network for video prediction using disentangled features," in *British Machine Vision Conference*, 2018.

[12] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[13] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Y. Wang, Z. Gao, M. Long, J. Wang, and S. Y. Philip, "PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5123–5132.

[15] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, S. Y. Philip, and M. Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208–2225, 2023.

[16] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International conference on learning representations*, 2019.

[17] J. Su, W. Byeon, J. Kossaifi, F. Huang, J. Kautz, and A. Anandkumar, "Convolutional tensor-train lstm for spatio-temporal learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 714–13 726, 2020.

[18] W. Saideni, D. Helbert, F. Courreges, and J. P. Cances, "A novel video prediction algorithm based on robust spatiotemporal convolutional long short-term memory (robust-st-convlstm)," in *Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 2*. Springer, 2022, pp. 193–204.

[19] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9154–9162.

[20] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "SimVP: Simpler yet better video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3170–3180.

[21] Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Strpm: A spatiotemporal residual predictive model for high-resolution video prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 946–13 955.

[22] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, "Video (language) modeling: a baseline for generative models of natural videos," *arXiv preprint arXiv:1412.6604*, 2014.

[23] C. Xu, P. Zhao, Y. Liu, J. Xu, V. S. S. S. Sheng, Z. Cui, X. Zhou, and H. Xiong, "Recurrent convolutional neural network for sequential recommendation," in *The world wide web conference*, 2019, pp. 3398–3404.

[24] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Advances in Neural Information Processing Systems*, 2018, pp. 517–526.

[25] T. Xue, J. Wu, K. Bouman, and B. Freeman, "Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 91–99.

[26] D. P. Kingma and M. Welling, "Stochastic gradient vb and the variational auto-encoder," in *Second International Conference on Learning Representations, ICLR*, vol. 19, 2014.

[27] M. Oliu, J. Selva, and S. Escalera, "Folded recurrent neural networks for future video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 716–731.

[28] H. Razali and B. Fernando, "A log-likelihood regularized kl divergence for video prediction with a 3d convolutional variational recurrent network," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 209–217.

[29] X. Ye and G.-A. Bilodeau, "Video prediction by efficient transformers," *Image and Vision Computing*, vol. 130, p. 104612, 2023.

[30] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, 2016, pp. 1558–1566.

[31] L. Castrejon, N. Ballas, and A. Courville, "Improved conditional vrnns for video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7608–7617.

[32] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in *International conference on machine learning*. PMLR, 2018, pp. 1174–1183.

[33] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[34] M. Babaeizadeh, C. Finn, D. Erhan, R. Campbell, and S. Levine, "Stochastic variational video prediction," in *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[35] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, 2016.

[36] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Advances in neural information processing systems*, 2016, pp. 64–72.

[37] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4463–4471.

[38] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[39] V. L. Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 474–11 484.

[40] H. Wu, Z. Yao, J. Wang, and M. Long, "Motionrnn: A flexible model for video prediction with spacetime-varying motions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 435–15 444.

[41] S. Lee, H. G. Kim, D. H. Choi, H.-I. Kim, and Y. M. Ro, "Video prediction recalling long-term motion context via memory alignment learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3054–3063.

[42] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[43] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[44] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in Neural Information Processing Systems*, 2019, pp. 14 837–14 847.

[45] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[47] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE transactions on Broadcasting*, vol. 54, no. 3, pp. 660–668, 2008.

[48] V. Sitzmann, M. Zollhöfer, and G. Wetzstein, "Scene representation networks: Continuous 3d-structure-aware neural scene representations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[50] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Fvd: A new metric for video generation," *ICLR*, 2019.

[51] S. Mrak *et al.*, "Reliability of objective picture quality measures," *Journal of Electrical Engineering*, vol. 55, no. 1-2, pp. 3–10, 2004.

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[53] X. Fei, L. Xiao, Y. Sun, and Z. Wei, "Perceptual image quality assessment based on structural similarity and visual masking," *Signal Processing: Image Communication*, vol. 27, no. 7, pp. 772–783, 2012.

[54] P. J. Kim, S. Kim, and J. Yoo, "Stream: Spatio-temporal evaluation and analysis metric for video generative models," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[55] B. Jin, Y. Hu, Q. Tang, J. Niu, Z. Shi, Y. Han, and X. Li, "Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4554–4563.

[56] X. Ye and G.-A. Bilodeau, "A unified model for continuous conditional video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3603–3612.

[57] S. Ning, M. Lan, Y. Li, C. Chen, Q. Chen, X. Chen, X. Han, and S. Cui, "Mimo is all you need: A strong multi-in-multi-out baseline for video prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1975–1983.

Mohammad Adiban is a PhD candidate in Machine Learning at the Norwegian University of Science and Technology (NTNU). He holds a Bachelor's degree in Computer Engineering and a Master's degree in Artificial Intelligence from the Sharif University of Technology, awarded in 2017. In 2022, he conducted research as a visiting scholar at Monash University in Australia. Additionally, Mohammad is a co-founder of the company Connect Me and Senior Data Scientist at Bluware company. His research focuses on statistical machine learning, signal processing, computer vision, speech processing, biomedical applications, and cyber security.

Kalin Stefanov is an ARC DECRA Fellow at the Faculty of Information Technology, Monash University, Melbourne, Australia. He received the MSc degree in Artificial Intelligence from the University of Amsterdam, Amsterdam, Netherlands and a PhD degree in Computer Science from KTH Royal Institute of Technology, Stockholm, Sweden. Prior to his current role, he was a Research Associate and Postdoctoral Research Scholar at the University of Southern California, Los Angeles, USA. His main research interests are machine learning, computer vision, and affective computing.

Sabato Marco Siniscalchi (Senior Member, IEEE) is a FULL Professor with the University of Palermo,Palermo, Italy, an Adjunct Professor with the Norwegian University of Science and Technology (NTNU), and an Affiliate Faculty with the Georgia Institute of Technology. He received his doctorate degree in computer engineering from the University of Palermo, Palermo, Italy, in 2006. In 2006, he was a Postdoctoral Fellow with Ga Tech. From 2007 to 2010, he joined NTNU, Norway, as a Research Scientist. From 2010 to 2023, he was an Assistant Professor, first, an Associate Professor, second, and a Full Professor, after, at Kore University. From 2017 to 2018, he was a Senior Speech Researcher with Siri Speech Group, Apple Inc., Cupertino CA, USA. He acted as an Associate Editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing, from 2015 to 2019. Prof. Siniscalchi was an Elected Member of the IEEE SLT Committee from 2019 to 2022 and was re-elected in 2024.

Giampiero Salvi (Senior Member, IEEE) is a Full Professor at the Department of Electronic Systems at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, and Associate Professor at KTH Royal Institute of Technology, Department of Electrical Engineering and Computer Science, Stockholm, Sweden. Prof. Salvi received the MSc degree in Electronic Engineering from Università la Sapienza, Rome, Italy and the PhD degree in Computer Science from KTH. He was a post-doctoral fellow at the Institute of Systems and Robotics, Lisbon, Portugal. He was a co-founder of the company SynFace AB, active between 2006 and 2016. His main interests are machine learning, speech technology, and cognitive systems.