Long-term Head Pose Forecasting Conditioned on the Gaze-guiding Prior

Shentong Mo*
Carnegie Mellon University
Pittsburgh, United States
shentonm@andrew.cmu.edu

Miao Xin†
Institute of Automation (CASIA), Chinese Academy of Sciences
Beijing, China
miao.xin@ia.ac.cn

Abstract

Forecasting head pose future states is a novel task in computer vision. Since future may have many possibilities, and the logical results are much more important than the impractical ones, the forecasting results for most of the scenarios should be not only diverse but also logically realistic. These requirements pose a real challenge to the current methods, which motivates us to seek for better head pose representation and methods to restrict the forecasting reasonably. In this paper, we adopt a spatial-temporal graph to model the interdependencies between the distribution of landmarks and head pose angles. Furthermore, we propose the conditional spatial-temporal variational graph autoencoder (CST-VGAE), a deep conditional generative model for learning restricted one-to-many mappings conditioned on the spatial-temporal graph input. Specifically, we improve the proposed CST-VGAE for the long-term head pose forecasting task in terms of several aspects. First, we introduce a gaze-guiding prior based on the physiology. Then we apply a temporal self-attention and self-supervised learning mechanism to learn the long-range dependencies on the gaze prior. To better model head poses structurally, we introduce a Gaussian Mixture Model (GMM), instead of a fixed Gaussian in the encoded latent space. Experiments demonstrate the effectiveness of the proposed method for the long-term head pose forecasting task. We achieve superior forecasting performance on the benchmark datasets compared to the existing methods.

1. Introduction

Head pose forecasting is to predict the 3D head angles in the future according to the current state. As a novel task in computer vision, it gains increasing attention recently due to widely needs in virtual reality [30], human-computer interaction [54], driving safety assistance [43], etc.

Head pose forecasting is more challenging than the classical estimation task [36, 59, 16, 58, 33]. Certain methods view it as a vector-based trajectory prediction problem, and employ regression models to predict the future state. However, as there are many possibilities in the future, learning simply one-to-one mapping may leads to amphibolous results. Recent methods utilize generative models [24, 29] to obtain future states, e.g., using generative adversarial networks (GAN) to generate future images then predict the 3D poses according to that. These methods can utilize unlabeled data and introduce the uncertainty into the generation. However, they heavily dependent on massive face images with high resolution. As a trade-off, we seek to utilize a better modality to represent the head pose.

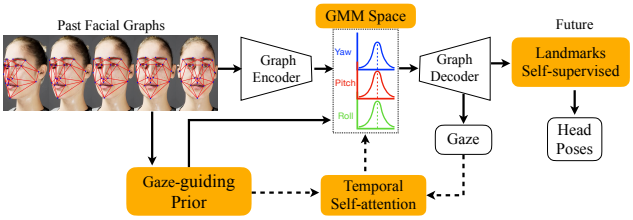

Figure 1. Pipeline of CST-VGAE for long-term head pose forecasting conditioned on the gaze-guiding prior.

Another challenge is how to achieve the forecasting results that are logical. The recent success of variational autoencoder (VAE) [24] shows evident advantages in learning latent representation. The VAE is capable of modeling *uncertainties* in structured one-to-many mapping by means of probabilistic inference. This nature is attractive for many tasks such as future forecasting [49]. However, despite allowing the diverse prediction, the states generated by VAE are inundated with lots of unrealistic or over-smoothing results [56], especially in long-term generation tasks. In real-world problems, the most possible and practical states are much more important than meaningless richness. A natural question is thus raised that ***how could we restrict the forecasting and achieve plausible results***. To this end, we need a constrained generation to guarantee the diverse and realistic long-term forecasting.

In this work, we propose to utilize the *graph* structure to represent the head pose, and frame the head pose forecasting as the spatial-temporal graph generation. We propose the *conditional spatial-temporal variational graph autoen-*

*This work was done during his intern at CASIA.
†Corresponding author.

coder (CST-VGAE), a deep generative model to learn restricted one-to-many mappings for graphs prediction. This model is capable of learning latent representation conditioned on the spatial-temporal graph-structured data. Our model is trained to maximize the conditional log-likelihood, and we formulate the variational learning objective of the CST-VGAE based on stochastic gradient variational Bayes (SGVB). The proposed CST-VGAE is capable of modeling a restricted one-to-many mapping to generate diverse and realistic results. As a result, this model is well-suited for long-term forecasting problems.

Specifically, we introduce a *gaze-guiding prior* into the head pose forecasting task, which is inspired by the physiological research. Furthermore, we apply a temporal *self-attention mechanism* to learn the long-range dependencies on the gaze prior. To better model head poses on three degree of freedom angles, we introduce a *Gaussian Mixture Model* (GMM), instead of a fixed Gaussian in our encoded latent space. Moreover, we leverage the *self-supervised learning* to restrain the long-term generation. As a result, the proposed CST-VGAE achieves superior performance as compared to most existing methods on this task. In summary, our contribution is as follows:

- We propose the CST-VGAE, a framework for self-supervised learning on spatial-temporal graph data, and introduce a physiology-based gaze-guiding prior for head pose long-term forecasting tasks.
- We apply a temporal self-attention mechanism and the self-supervised learning method to learn the long-range dependencies on the gaze prior.
- We introduce a Gaussian Mixture Model (GMM), instead of a fixed Gaussian in our gaze prior to better model head poses on three degree of freedom angles.
- We achieve superior long-term head pose forecasting performance on two benchmark datasets as compared to existing models. We also verify the method's cross-task generalization on the body pose forecasting task.

2. Related work

Head pose forecasting. Head pose forecasting is to predict the future 3D head pose vector according to the current 2D information. A closely related task is the *head pose estimation*, which recognizes the current head pose by means of facial landmarks[5, 28] or images directly [58, 41]. Despite the latter are more compact in model size, these methods call for massive data than landmark-based methods [31], otherwise they suffer from over-fitting since head pose regression is a fine-grained task. Comparing with head pose estimation, the head pose forecasting is more challenging since the future may have many possibilities while we care more about the situations of higher importance under certain conditions, e.g., distractions before the traffic accidents. Although this is a novel problem in computer vision, future prediction *w.r.t.* body poses is studied in recent years [52, 50]. Most of these methods leverage generative models to learn one-to-many mappings.

Deep generative models. Along with the recent breakthroughs in unsupervised learning methods, there has been progress in deep generative models, such as the generative adversarial network (GAN) [15] and variational autoencoder (VAE) [24]. Recently, the advances in inference and learning algorithms for various deep generative models [44, 62] significantly enhanced this line of research. Conditional variational autoencoder (CVAE) [46] is proposed to model the distribution of high dimensional output space as a generative model conditioned on the input observation. Variational graph autoencoder (VGAE) [25] is proposed for unsupervised learning on non-Euclidean, graph-structured data based on the VAE. In this work, we propose the CST-VGAE for modeling the prior on Gaussian latent representations given graph-structured input data. Perhaps the most similar prior work to our approach is [32] in which the condition is to generate the connection of the nodes, while our method is to restrict the generation of the nodes' spatial features. Hence, our method falls into the category of spatial-temporal graph generation [55].

Self-attention. Self-attention stems from the machine translation [48] and the natural language processing [9, 60]. In recent years, self-attention has been also explored in computer vision tasks, such as image recognition [67, 2], image captioning [8, 61], and image synthesis [4, 38, 65]. Han Zhang et al. [65] propose Self-Attention Generative Adversarial Network (SAGAN) to focus on detailed features of the distant portions of generated images. Another related work [47] applies a self-attention module to capture the long-range dependencies of point cloud object 3D shapes for diverse and realistic generation. However, little work in the literature adopts a self-attention module in long-term forecasting problems. To the best of our knowledge, this work is the first to propose a self-attention module in the long-term head pose forecasting task.

Self-supervised learning. Self-supervised learning is a subset of unsupervised learning methods. Self-supervised learning approaches for deep learning have been shown to be effective in various domains, such as audio and video analysis [26], image decomposition [21], point clouds [42] and human pose reconstruction [40]. On long-term forecasting tasks, approaches have largely focused on utilizing either variational autoencoders [49] or autoencoders combined with long-short-term-memory (LSTMs) [19]. Self-supervised learning is supervised learning without human-annotated labels. In this work, our proposed model is applied to forecast the long-term (next one-second) head pose in a video given past frames. This can be regarded as an instance of self-supervised learning or temporally supervised learning, where supervision comes from future input data.

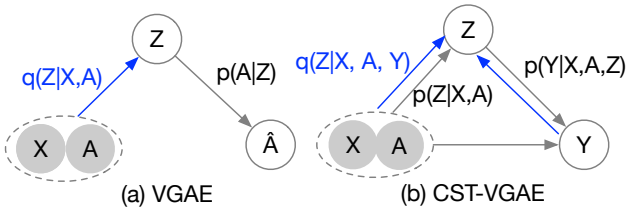

Figure 2. Illustration of (a) VGAE and (b) CST-VGAE. Blue arrow lines represent the inference process $q(\cdot)$. Gray arrow lines denote the generation process $p(\cdot)$.

3. Conditional Spatial-temporal VGAE

Preliminary: Variational Graph Auto-encoder. The variational graph auto-encoder (VGAE) [25] is a framework for unsupervised learning on graph-structured data based on the variational autoencoder (VAE) with the Gaussian latent variables. Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n nodes, a set of latent variable $\mathbf{z}_i \in R^m$ is generated from the prior distribution $p(\mathbf{Z})$, and the adjacency matrix $\mathbf{A}$ of $\mathcal{G}$ is generated by the generative distribution $p(\mathbf{A}|\mathbf{Z})$ conditioned on latent variables $\mathbf{z}_i$: $\mathbf{z}_i \sim p(\mathbf{Z})$, where $\mathbf{z}_i$ is summarized in the matrix $\mathbf{Z} \in \mathbf{R}^{n\times m}$. Figure 2(a) sketches this process. To address intractable posterior inference problems, the VGAE optimizes the variational lower bound L *w.r.t.* the variational parameters. The variational lower bound of the VGAE is written as follows:

$$L_{VGAE} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{A})}[\log p(\mathbf{A}|\mathbf{Z})] - KL[q(\mathbf{Z}|\mathbf{X},\mathbf{A})\|p(\mathbf{Z})], \tag{1}$$

where $\mathbf{X} \in \mathbf{R}^{n\times d}$ denotes the node features matrix. $KL[q(\cdot)\|p(\cdot)]$ is the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$. $p(\mathbf{A}|\mathbf{Z})$ represents the generative model. $q(\mathbf{Z}|\mathbf{X},\mathbf{A})$ denotes the inference model. $\mathbf{Z} = g(\mathbf{X},\mathbf{A},\epsilon)$, $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$.

CST-VGAE. The aforementioned VGAE simply models a one-to-many mapping to predict diverse results. However, in real-world long-term forecasting tasks, the models are required to generate results that satisfied certain temporal constrains; otherwise the outputs do not conform to realism. This motivates us to structure the latent space, achieving by modeling the conditional distribution. Therefore, in this work, we propose a *conditional spatial-temporal variational graph auto-encoder* (CST-VGAE) model for learning one-to-many mapping conditioned on temporal constraints.

As illustrated in Figure 2(b), for the spatial-temporal graph [55] (an attributed graph where the node attributes change dynamically over time), the conditional generative process of the CST-VGAE is as follows: for given input $\mathcal{I}_{\mathbf{X},\mathbf{A}}$ with node features matrix $\mathbf{X}$ and an adjacency matrix $\mathbf{A}$, a set of latent variables $\mathbf{Z}$ is sampled from the prior distribution $p(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}})$. In contrast to the VGAE, the output $\mathcal{Y}_{\mathbf{X},\mathbf{A}}$ with node features matrix $\mathbf{X}$ and an adjacency matrix $\mathbf{A}$ is generated from the distribution $p(\mathcal{Y}_{\mathbf{X},\mathbf{A}}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathbf{Z})$. Notice that we follow a typical form of the spatial-temporal graph [55], meaning that the graph structure is fixed.

The CST-VGAE is trained to maximize the *conditional log-likelihood*. Often the objective function is intractable, and thus we apply the reparameterization trick to train the model. The variational lower bound of the model is written as follows:

$$\begin{aligned}\log p(\mathcal{Y}_{\mathbf{X},\mathbf{A}}|\mathcal{I}_{\mathbf{X},\mathbf{A}}) \geq & -KL[q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathcal{Y}_{\mathbf{X},\mathbf{A}})\|p(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}})] \\ & + \mathbb{E}_{q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathcal{Y}_{\mathbf{X},\mathbf{A}})}[\log p(\mathcal{Y}_{\mathbf{X},\mathbf{A}}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathbf{Z})],\end{aligned} \tag{2}$$

and the variational lower bound $L_{CST-VGAE}$ is written as:

$$\begin{aligned}L_{CST-VGAE} = & \mathbb{E}_{q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathcal{Y}_{\mathbf{X},\mathbf{A}})}[\log p(\mathcal{Y}_{\mathbf{X},\mathbf{A}}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathbf{Z})] \\ & - KL[q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathcal{Y}_{\mathbf{X},\mathbf{A}})\|p(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}})],\end{aligned} \tag{3}$$

where $p(\mathcal{Y}_{\mathbf{X},\mathbf{A}}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathbf{Z})$ represents the conditional generative model. $p(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}})$ represents the prior of the latent variable $\mathbf{Z}$, $\mathbf{Z} = g(\mathcal{I}_{\mathbf{X},\mathbf{A}},\epsilon)$, $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. Note that the inference distribution $q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}},\mathcal{Y}_{\mathbf{X},\mathbf{A}})$ is reparametrized with a deterministic, differentiable function $g(\cdot,\cdot)$, whose arguments are input data $\mathcal{I}_{\mathbf{X},\mathbf{A}}$, and the noise variable ϵ.

4. CST-VGAE for long-term head pose forecasting

We employ the CST-VGAE to address the long-term head pose forecasting problem. Our experiments (see Section 6.2) suggest that the head pose forecasting via the image generation is impractical, especially when the training data is insufficient. By contrast, the approach that takes advantage of the sparse and low-dimensional facial landmarks as well as the topology demonstrates more promising results. In our work, we select 19 facial landmarks that invariable to facial expressions to form the landmark graph (see Figure 3), and the landmark coordinates are viewed as the node feature $\mathbf{x}_i$.

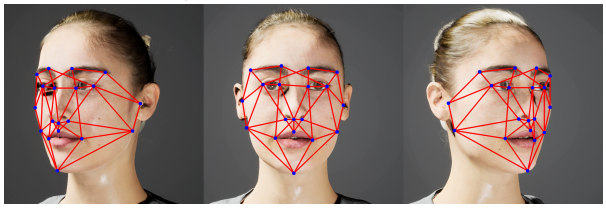

Figure 3. Visualization of the landmark graph.

Overview. In general, our model is composed of an encoder and a decoder (see Figure 4), both of which are constructed by the ST-GCNs [57]. The **encoder** takes facial graphs as an input, and generates the latent vector Z under certain conditions. The **decoder** samples from Z to generate landmarks and regresses the 3D head pose vectors. The head pose forecasting task implies important prior knowledge that can help us to improve the vanilla CST-VGAE effectively on this task. So we have a series of designs in the improved model. 1) To handle the long-term forecasting issue, we need the temporal restriction. So we introduce

the self-attention mechanism to the vanilla CST-VGAE. 2) To improve the long-term forecasting performance, we also adopt the landmarks predicted from the future as temporally self-supervised learning. 3) For better diversity and interpretability in generation results, we seek to replace the fixed Gaussian with the Mixture Gaussian prior. 4) In addition to that, we need a proper condition to restrain the generation. Inspired by physiology researches, we introduce the gaze-guiding prior in this work.

4.1. Gaze-guiding prior

In physiology, one interesting phenomenon [12, 35] is that the human gaze usually predates the head pose in the real world, i.e., the gaze direction at $t-j$ indicates the future head pose at $t+l$. This phenomenon suggests that the gaze direction may act as the prior condition. To confirm this observation, we evaluate the correlation between the gaze and the head pose. We calculate the *Pearson correlation coefficient* of gaze vector (θ, ϕ) and head pose measurements including $(yaw, pitch, roll)$ angles, where the gaze predates the head pose with time-step 5 frames. We utilize a gaze tracking model [22] to estimate gaze directions of corresponding frames on the BIWI Kinect Head Pose database [11]. The output (θ, ϕ) of the gaze tracking is the expected gaze direction in spherical coordinates. As shown in Figure 5 (Left), almost each Pearson correlation coefficient between (θ, ϕ) and $(yaw, pitch, roll)$ is larger than 0.7, confirming that the head pose is strongly related to the gaze direction happening before it. We conduct extensive experiments (see Section 6.1) to decide the optimal time-step interval between the gaze direction and the head pose. Thus, we introduce the essential **gaze-guiding prior** as the condition of the proposed model in the head pose long-term forecasting task.

4.2. Self-attention temporal condition

To explicitly model the contribution of each current gaze prior, we introduce a self-attention mechanism to our gaze prior that is composed of a fixed number (T) of consecutive frames. The fixed number of frames is 5 in our experiments.

In order to capture long-range dependencies on the gaze prior, we adopt a recurrent self-attention module between the gaze prior c_i and the predicted gaze g_i at the recurrent iteration i, as shown in Figure 4 (dotted arrows). This means that the predicted future gaze g_i will be fed back into our self-attention module (see Algorithm 1) to form a new gaze prior c_{i+1} at recurrent iteration $i+1$. As a result, the proposed CST-GVAE model can be used to forecast long-term head poses recurrently. To the best of our knowledge, we are the first to propose a **self-attention temporal condition** for long-term head pose forecasting tasks.

As Algorithm 1 shows, we apply a multi-head attention module in our self-attention temporal condition c_i, since we need different attention heads for focusing on different aspects of the temporal condition c_i. $\#head$ represents the number of heads in Algorithm 1. First, our gaze prior of dimension size 4 is projected onto a high dimension space (D). After projection, the gaze prior c_i at recurrent iteration i is used as the query q, and the predicted gaze g_i at this iteration is for the key k and the value v. Then the scaled dot-product attention is applied to q_h, k_h, v_h in each head. Finally, the multi-head attention results are concatenated together and a linear layer is applied to obtain the new gaze prior c_{i+1}.

Algorithm 1: Self-attention condition module

Input: c_i: gaze condition at iteration i
g_i: gaze forecasted at iteration i
Output: c_{i+1}: gaze condition at iteration $i+1$

$c_i, c_{i+1}, g_i \in \mathbb{R}^{T\times 4}$
$W_h^q, W_h^k, W_h^v \in \mathbb{R}^{4\times D}, W^o \in \mathbb{R}^{(\#head\times D)\times 4}$
Function `Self-attention(`c_i, g_i`)`**:**
  while $h < \#head$ **do**
    $q_h \leftarrow c_i W_h^q$
    $k_h, v_h \leftarrow g_i W_h^k, g_i W_h^v$
    attention_$h \leftarrow \text{softmax}\left(\frac{q_h k_h^\top}{\sqrt{D}}\right)\cdot v_h$
  attention $\leftarrow$ Concat{attention_h}
  $c_{i+1} \leftarrow$ attention$\cdot W^o$
  return c_{i+1}

4.3. Mixture distribution prior

The head pose states consist of three degree of freedom angles (yaw, pitch, roll). This prominent knowledge motivates us to introduce a **mixture of Gaussian distribution**, instead of a single Gaussian, to model our gaze prior in the proposed CST-GVAE model. As shown in Figure 4, the proposed model can explicitly encode the latent space around three components corresponding to yaw (μ_y, σ_y), pitch (μ_p, σ_p), roll (μ_r, σ_r), respectively, and combine all components to create a GMM latent space for head pose forecasting.

Note that our proposed GMM prior is distinct from one work [53] in the literature. Specifically, there are two main differences between ours and their methods. The first thing is that they create a GMM latent space in the vanilla CVAE for diverse image description, while we propose the GMM prior for forecasting graph node features. The second one is that we use different standard deviation σ for all components, but they use the same σ for all components.

4.4. Self-supervised training

In order to forecast the long-term head pose in a video given past frames, we adopt the landmarks predicted from the future as temporally supervised learning, where it is an instance of **self-supervised learning** and supervision comes from the future input data [64]. With the landmark

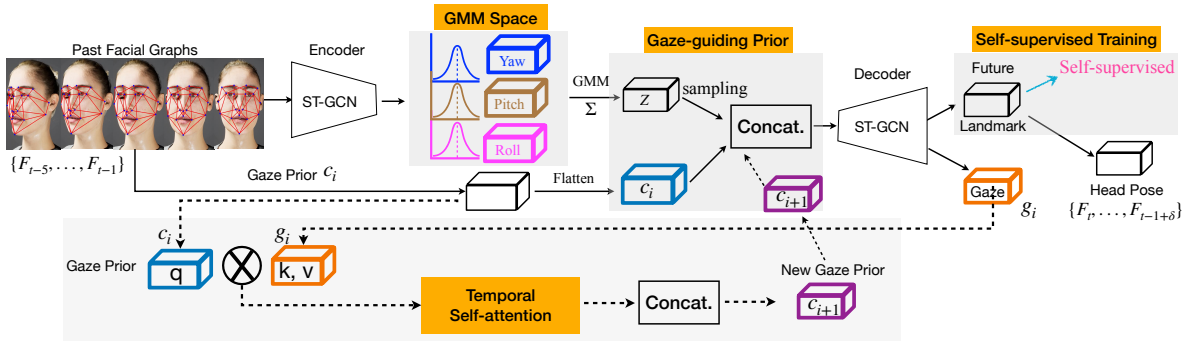

Figure 4. Overview of the **CST-VGAE** on head pose long-term forecasting tasks. The **solid** arrows denotes the current iteration i, and the **dotted** arrows represents the next iteration $i+1$. At iteration i, the input composed of the facial landmark graph at $\{F_{t-5}, ..., F_{t-1}\}$ are fed into the ST-GCN encoder to generate a mixture Gaussian latent embeddings Z that are concatenated with the gaze prior c_i to do sampling. The ST-GCN decoder outputs 19 landmarks as self-supervised information to predict head poses at $\{F_t, ..., F_{t-1+\delta}\}$. Here δ is the time step. The decoder also predicts gazes g_i for combining with c_i to feed back into temporal self-attention module to generate a gaze prior c_{i+1} for next iteration $i+1$. Therefore, our CST-VGAE model forecasts the long-term head poses in a recurrent way.

based self-supervised learning, our CST-VGAE intends to forecast the head pose at longer steps in the future.

5. Implementation details

Data structure. In our experiment, we implement FAN [5], a state-of-the-art face alignment method to extract 19 facial landmarks from each past frame to form the landmark graph (see Figure 3). For the gaze prior, we utilize OpenPose [7, 45] to extract two pupil landmarks. In the training process, we take a graph (composed of 19 facial landmarks and 2 pupils) from five past frames to feed into the encoder, forming an input of shape $(B, 5, 21, 2)$. B denotes the batch size and 5 is the number of input frames in our experiments. The dimensionality of latent embeddings is fixed to 32 for all models (see Section 6.1 for the detailed discussion).

For the gaze prior, we also apply a self-attention module to the 2 pupils from 5 past frames for allowing a temporal embedding of shape $(B, 2 \times 2 \times 5)$. Then we concatenate the self-attention gaze prior with latent embeddings from the encoder. The decoder predicts 19 facial landmarks and 2 pupils as self-supervised information to train our model. Finally, our pose decoder will output a fixed number (T) of forecasted frames with three head pose vectors, forming shape $(B, 3 \times T)$. In the training process, we find that the model suffers from KL divergence vanishing after limited iterations, leading to generate unstable latent representations. Motivated by beta-VAE [17, 6] and other KL divergence annealing methods [14, 3], we introduce a coefficient γ to constrain the KL divergence, thus the objective function $L^*_{CST-VGAE}$ is written as follows:

$$L^*_{CST-VGAE} = \mathbb{E}_{q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}}, \mathcal{Y}_{\mathbf{X},\mathbf{A}})}[\log p(\mathcal{Y}_{\mathbf{X},\mathbf{A}}|\mathcal{I}_{\mathbf{X},\mathbf{A}}, \mathbf{Z})] - \gamma \times KL[q(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}}, \mathcal{Y}_{\mathbf{X},\mathbf{A}})||p(\mathbf{Z}|\mathcal{I}_{\mathbf{X},\mathbf{A}})], \quad (4)$$

where γ represents the weight of the KL divergence term.

Our implementation [1] is built on PyTorch [39]. We use Adam [23] for the adaptive learning rate scheduling algorithm of gradient descent optimization. The training is executed on NVIDIA GeForce RTX 2080 Ti GPU for 31.2 minutes. We set learning rate = $5e^{-4}$, batch size = 1024, epoch = 300. $T = 5$ in our experiments.

Inference. Our model forecasts long-term head poses in a recurrent way. In each recurrent iteration, we use a fixed number of frames ($T = 5$) of gaze prior and forecast a fixed number of frames of self-supervised landmarks. We feed the forecasted future gaze g_i at recurrent iteration i back into our self-attention module to form a new gaze prior c_{i+1} at recurrent iteration $i+1$. In this way, the CST-VGAE model can capture the long-range dependencies between these gazes for guiding head poses in the future. As a result, the proposed CST-VGAE model can be used to forecast long-term head poses in a recycled fashion. We test 30,000 instances from BIWI and UPNA datasets for 5 runs on one 2080 Ti GPU. The average inference time per instance is 0.1 seconds.

6. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of the CST-VGAE. First, we implement three preliminary experiments w.r.t. the hyper-parameter γ, the gaze-guiding prior, and the dimensionality of the latent embeddings. Then, we compare our proposed CST-VGAE with the state-of-the-arts on BIWI [11] and UPNA [1] datasets from quantitative and visualization results. Finally, we perform detailed ablation studies to demonstrate the effectiveness of each module in CST-VGAE and the cross task generalization of our model.

Dataset. The BIWI Kinect Head Pose Database [11] contains 24 videos of 20 subjects (a total of roughly 15,000 frames) in the controlled environment. We compare existing methods based on the Mean Absolute Error (MAE) of

[1] https://sites.google.com/view/cst-vgae/

the forecasted head pose. In this protocol, we use 70% of videos (16 videos) in the BIWI dataset for training, and the others (8 videos) for testing. In our experiments, we only consider frames whose head pose rotation angles are within the range of $[-90°, +90°]$. The UPNA Head Pose Database [1] contains 10 subjects, where each user has 12 videos and 300 frames per video. This dataset is used for validate the generalizability of our CST-VGAE.

6.1. Preliminary experiments

First, we conduct the preliminary experiments for exploring the hyper-parameters in the proposed CST-VGAE.

Hyper-parameter γ. We set γ as a constant in the range of $(0, 1]$ and adopt state-of-the-art KL divergence annealing methods [17, 14, 3] to adjust the weights of the KL divergence term. We report the comparison results in Table 1. When we set γ as a constant $5e^{-2}$, our model performs well on both the MAE and KL divergence. By contrast, the previous annealing methods have relatively poor results on reducing the MAE in this case.

Optimal time-step interval. Another hyper-parameter is the *optimal time-step interval* between the gaze-guiding prior and the head pose. To determine this time interval, we set 14 different frame intervals between the head pose and the gaze direction to test on the BIWI dataset. Figure 5 (Right) describes this trend that the *Pearson correlation coefficient* increases first and decrease later with frame intervals ascending. When the frame interval is approximately from 5 to 7, the correlation between the head pose and the gaze direction arrives at the peak 0.861. Accordingly, we choose 5 as the frame interval.

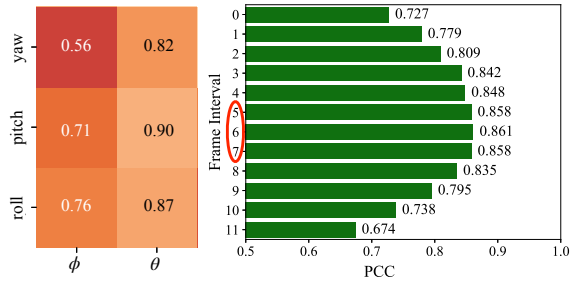

Figure 5. Left: Pearson correlation coefficient heat-map of the predated gaze-guiding prior and head pose. Right: distribution of Pearson correlation coefficient for 12 different frame intervals (0-11) between the gaze-guiding prior and the head pose. Red eclipse part denotes the optimal frame interval.

Dimensionality of the latent embeddings. We explore six custom dimensionality sizes (16, 32, 64, 128, 256, 512) to forecast long-term head poses $\{F_t, F_{t+1}, ...F_{t+29}\}$ for comparison. Table 2 describes the results of this experiment with the proposed CST-VGAE model. When the dimensionality size is equal to 32, the MAE metric performs the best. The larger dimensionality size is, the KL divergence becomes larger. Moreover, neither the MAE performance nor the model size is acceptable.

Table 1. Comparison results of different annealing methods for adjusting the hyper-parameter γ.

Method	Yaw	Pitch	Roll	MAE	KL
step [3]	2.35	2.64	7.02	4.00	2.45
linear [3]	2.73	2.39	6.96	4.03	8e-5
logistic [3]	2.44	2.57	6.63	3.88	1e-4
cyclical [14]	2.61	2.62	6.46	3.90	1e-3
$\gamma = 10$ [17]	2.86	2.94	7.23	4.34	1e-5
$\gamma = 1$	2.76	2.7	7.03	4.16	2e-4
$\gamma = 5e^{-1}$	2.67	2.58	6.73	3.99	5e-4
$\gamma = 1e^{-1}$	2.3	2.23	6.73	3.75	4.29
$\gamma = 5e^{-2}$	**2.45**	**2.26**	**6.23**	**3.65**	**7.46**
$\gamma = 1e^{-2}$	2.87	2.65	6.45	3.99	22.26
$\gamma = 1e^{-3}$	3.32	2.97	6.76	4.35	84.92
$\gamma = 1e^{-4}$	2.4	2.28	6.15	3.61	756.69
$\gamma = 1e^{-5}$	2.68	2.27	5.85	3.60	2674.64

Table 2. Comparison results of different latent embedding sizes ("Dim" for dimension size, "M-Size" for model size).

Dim	Yaw	Pitch	Roll	MAE	KL	M-Size(MB)
16	3.19	2.91	6.71	4.27	22.68	2.38
32	**2.45**	**2.26**	**6.23**	**3.65**	**7.46**	3.47
64	2.51	2.34	6.39	3.75	22.99	8.43
128	2.84	2.61	6.32	3.92	23.55	27.9
256	2.7	2.63	6.27	3.86	25.79	105
512	4.71	5.16	7.51	5.79	26.28	411

6.2. Comparison with the state-of-the-arts

In this section, we implement extensive experiments to compare the proposed CST-VGAE model with the state-of-the-arts. There are three main categories for comparison. The *first* category is deterministic prediction methods using facial landmarks as input to regress yaw, pitch, roll directly, including Linear Regression, Vanilla LSTM [18], GAZE-LSTM, ERD [13], and SRNN [20],. Vanilla-LSTM is a standard LSTM encoder-decoder model, and GAZE-LSTM incorporates the gaze prior. The *second* category is stochastic and diversity prediction methods utilizing face images as input. We implement the VAE [24], GMVAE [10], and CycleGAN [29] to generate long-term faces first and then estimate head poses from these forecasted faces using HopeNet [41]. The *third* category falls into methods using facial landmark graphs as input. We choose the GAE [25], VGAE [25] for forecasting landmarks directly using the facial landmark graph as input. For a fair comparison, we use a comparable pose estimator [41] to estimate head poses from our forecasted landmarks. Our goal is to predict head poses given past frames at the next δ steps, *i.e.*, 1, 10, 20, 25, and 30. We evaluate those baselines and our CST-VGAE on the BIWI for long-term head pose forecasting by quantitive and visualization analysis. We also evaluate on the UPNA for cross-dataset generalizability validation. We report the quantitive results in Table 3.

First Category. Although GAZE-LSTM achieves lowest MAE_δ among those baselines, its error score is still larger than our CST-VGAE's by a large margin, which validates the advantage of our method in the long-term forecasting. We also input history head poses to Linear Regression directly, but its performance is unsatisfactory.

Table 3. Comparison results with the state-of-the-arts on BIWI and UPNA dataset. MAE_δ denotes the mean absolute error for forecasting future head poses at next δ steps. Bold and underline numbers denote the first and second place.

Methods	Landmark	Image	Graph	BIWI					UPNA				
				MAE_1	MAE_{10}	MAE_{20}	MAE_{25}	MAE_{30}	MAE_1	MAE_{10}	MAE_{20}	MAE_{25}	MAE_{30}
Linear Regression	✓			5.56	7.55	9.42	10.86	12.54	6.23	9.85	13.23	16.93	19.26
Vanilla-LSTM	✓			3.76	5.84	7.57	8.93	9.96	4.82	7.54	10.68	12.89	14.16
GAZE-LSTM	✓			3.32	5.42	<u>6.89</u>	<u>8.04</u>	<u>9.12</u>	4.32	6.71	8.54	9.86	11.25
ERD	✓			3.67	5.63	7.24	8.85	9.78	4.67	7.43	9.87	12.65	13.45
SRNN	✓			3.56	5.56	7.13	8.65	9.64	4.53	7.32	9.75	11.54	12.85
VAE		✓		17.99	18.06	18.15	18.17	18.00	19.02	21.13	22.04	23.05	25.16
CycleGAN		✓		13.79	18.48	18.79	18.09	19.04	16.45	21.89	24.43	25.78	27.02
GMVAE		✓		<u>3.25</u>	<u>5.31</u>	7.01	8.43	9.42	4.23	7.21	9.53	10.62	11.34
GAE			✓	4.23	5.81	8.39	8.77	9.78	4.53	6.82	8.84	9.82	11.43
VGAE			✓	3.56	5.71	7.90	8.50	9.21	<u>4.17</u>	<u>6.57</u>	<u>8.45</u>	<u>9.65</u>	<u>11.02</u>
CST-VGAE			✓	**3.08**	**4.16**	**4.82**	**5.80**	**6.27**	**3.45**	**4.56**	**5.46**	**6.15**	**6.56**

Second Category. For the second category, the CST-VGAE model outperforms VAE and CycleGAN by a large margin in terms of MAE. This is because the human faces are diverse and high-dimensional. It is impractical to forecast long-term head poses using the pixel-to-pixel generation.

Third Category. For the third comparison category, we find that the GAE performs worse than the VAGE. This is because the GAE does not incorporate any sampling in the generation process, making it ineffective for the head pose forecasting problem. By introducing the latent space for sampling, the VGAE performs better than the GAE. Moreover, both the performance of GAE and VGAE beat the VAE and CycleGAN. Therefore, facial landmark graphs are better suitable as input for long-term head pose forecasting. More importantly, with the introduction of the gaze-guiding prior, our proposed CST-VGAE model outperforms the aforementioned four methods in terms of MAE.

Cross-dataset Evaluation. We evaluate the model trained on BIWI dataset on videos of 10 subjects from UPNA dataset in Table 3. Our CST-VGAE achieves the lowest score in terms of all metrics, compared to baselines. This further shows that our method performs promising generalization to other real datasets.

Visualization. We visualize the head pose results on the original RGB images for GAE, VGAE and our CST-VGAE model. Instead, we directly show the generated face images by VAE and CycleGAN. As shown in Figure 6, we draw the results of 5 frames $\{F_t, F_{t+4}, F_{t+9}, F_{t+19}, F_{t+29}\}$ from forecasted 30 frames $\{F_t, ..., F_{t+29}\}$ for one subject. From Figure 6, the generated face images by VAE and CycleGAN are coarse and blur, making it hard to detect landmarks. We use HopeNet [41] to estimate head poses without keypoints. As a result, neither two ways are suitable for long-term forecasting compared to models using facial landmark graphs as input. By comparison, our CST-VGAE model achieves competitive long-term forecasting performance.

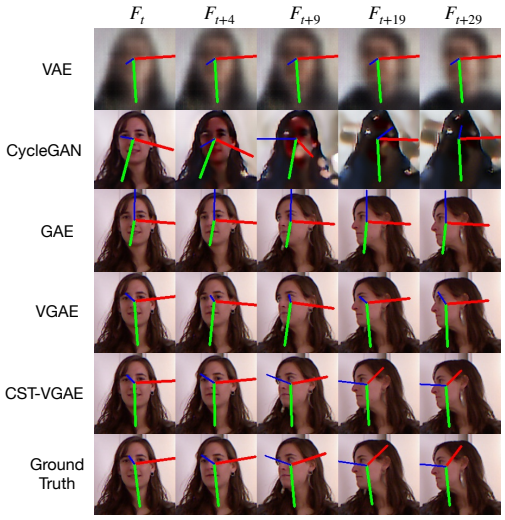

Figure 6. Visualization of the generated head pose results of one subject from the forecasted 30 frames. The six rows from top to bottom represent the **VAE**, **CycleGAN**, **GAE**, **VGAE**, **CST-VGAE**, and **Ground-Truth**, respectively. The **blue**, **green**, **red** lines denote the direction the subject is facing, the downward direction, and the side, respectively. Best viewed on screen.

6.3. Ablation study

Gaze-guiding prior. Table 4 compares the long-term forecasting performance of CST-VGAE without and with the gaze-guiding prior in terms of forecasting $20, 25, 30$ frames of head poses. The CST-VGAE with the gaze guiding prior outperforms the VGAE without the gaze prior in terms of three different numbers frame numbers of the forecasted head poses. This means that our gaze prior plays a guiding role in forecasting the long-term head poses.

Self-attention. To demonstrate the effectiveness of the self-attention temporal condition module in CST-VGAE with the gaze-guiding prior, Table 4 compares the performance of the proposed CST-VGAE with and without the self-attention module in terms of forecasting $20, 25, 30$ frames of head poses. The results show that the CST-VGAE with the self-attention module achieves better performance than that without the attention module, which confirms the effectiveness of our self-attention module in capturing the long-range dependencies on the gaze prior recurrently.

Mixture distribution prior. Based on the CST-VGAE with gaze-guiding prior and self-attention module, we also evaluate the performance of the proposed CST-VGAE model with single Gaussian prior and mixture Gaussian

Table 4. Ablation study performance on BIWI dataset. GP, TSA, MG, and SSL denote the gaze-guiding prior, temporal self-attention, mixture Gaussian, and self-supervised learning, respectively.

GP	TSA	MG	SSL	δ=20				δ=25				δ=30			
				Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
				9.35	8.62	5.74	7.90	9.55	9.63	6.31	8.50	10.58	10.61	6.43	9.21
✓				5.13	5.27	5.34	5.25(↓**2.65**)	6.36	6.56	5.97	6.30(↓**2.20**)	7.05	6.86	6.23	6.71(↓**2.50**)
✓	✓			4.76	4.96	5.14	4.95(↓**0.30**)	6.03	6.26	5.66	5.98(↓**0.32**)	6.85	6.45	6.03	6.44(↓**0.27**)
✓	✓	✓		4.65	4.78	5.04	4.82(↓**0.13**)	5.86	6.06	5.47	5.80(↓**0.18**)	6.65	6.28	5.87	6.27(↓**0.17**)
✓	✓	✓	✓	4.41	4.53	4.72	4.55(↓**0.27**)	5.42	5.62	5.11	5.38(↓**0.42**)	6.31	6.02	5.43	5.92(↓**0.35**)

prior in terms of forecasting $20, 25, 30$ frames of head poses. Table 4 reports the detailed comparison performance results. The results show that incorporating mixture Gaussian prior further improves our model.

Self-supervised Learning. In order to validate the effectiveness of the landmark based self-supervised learning in the long-term forecasting, we evaluate the performance of our CST-VGAE with and without the self-supervised learning in Table 4. We can observe an obvious performance improvement in terms of $20, 25, 30$ frames of head poses forecasting. This infers the significance of the landmark based self-supervised learning in the proposed CST-VGAE.

Table 5. Error analysis w.r.t. landmark & gaze errors.

Detector	MLE	MGE	MAE_{20}	MAE_{25}	MAE_{30}
DAN [27]	4.30	-	4.87	5.87	6.35
FAN[5]	4.06	-	4.82	**5.80**	6.27
HRNetV2 [51]	**3.85**	-	**4.79**	5.82	**6.22**
GazeNet [66]	-	5.50	4.72	5.65	6.14
GazeML [37]	-	**4.50**	**4.65**	**5.57**	**6.08**
OpenPose [7]	-	6.52	4.82	5.80	6.27

6.4. Error analysis and failure cases

We implement three SOTA facial landmark estimator and three gaze estimation methods to analyze how much the Mean Landmark estimation Error (MLE) and Mean Gaze estimation Error (MGE) affect the final forecasting accuracy in Table 5. Even though DAN's MLE is 0.24 larger than FAN's MLE, the differences between their MAE_δs are 0.05, 0.07, 0.08. This validates the robustness of our method to facial landmarks. However, the final accuracy becomes higher with the decrease of MGE, which further shows the importance of the proposed gaze-guiding prior. We show some failure cases in Figure 7. Our model sometimes misses the cases where the face and the gaze stay in the middle for a period longer than 3 seconds.

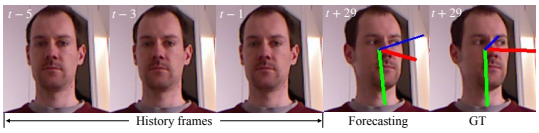

Figure 7. Failure cases analysis.

6.5. Cross task evaluation

In order to evaluate the generalizability of our proposed CST-VGAE to the field of body pose forecasting, we implement Linear Regression, Vanilla-LSTM, GAE, VAGE, and take body landmarks as input for fair comparison. We also compare our CST-VGAE with other previous work focusing on human pose forecasting, including ERD [13], LSTM-3LR [13], SRNN [20], Residual [34], GMVAE [10] and DLow [63]. Specifically, we evaluate those methods on the Kinetics-skeleton dataset [57] to predict the future 18 body landmarks given past frames at the next δ steps, *i.e.* 1, 10, 20, 25, and 30, and use the mean absolute error MAE_δ for evaluation. In our CST-VGAE, we incorporate the motion fields of past frames extracted by a motion encoder [56], as our prior. The quantitative results are reported in Table 6. From Table 6, our CST-VGAE performs comparably to DLow in terms of MAE_{20} and MAE_{25}, and achieves better results in terms of other metrics. This implies an excellent generalizability performance of our CST-VGAE to the body pose forecasting task.

Table 6. Comparison of performance of body pose forecasting on Kinetics-skeleton dataset.

Method	MAE_1	MAE_{10}	MAE_{20}	MAE_{25}	MAE_{30}
Linear Regression	6.86	9.45	11.92	13.66	16.43
Vanilla-LSTM	5.86	7.23	8.87	10.04	12.16
GAE	6.31	8.42	9.95	11.37	13.58
VGAE	5.24	6.87	8.98	10.14	12.67
CST-VGAE(ours)	**5.01**	**6.23**	7.85	9.36	**10.32**
DLow	5.11	6.25	**7.83**	**9.22**	10.48
GMVAE	5.52	6.68	8.35	9.46	11.42
Residual	5.64	6.78	8.48	9.53	11.53
SRNN	5.67	6.85	8.56	9.62	11.68
LSTM-3LR	5.71	7.09	8.69	9.76	11.82
ERD	5.82	7.18	8.76	9.87	11.94

7. Conclusions

In this work, we propose the *conditional spatial-temporal variational graph auto-encoder* for learning constrained one-to-many mappings conditioned on spatial-temporal graph input. We introduce the gaze-guiding prior as the condition in the CST-VGAE model for long-term head pose forecasting problems. We apply the temporal self-attention and self-supervised mechanism to learn the long-range dependencies on the gaze prior. We use a mixture Gaussian prior to explicitly encode the latent space around yaw, pitch, roll component, respectively. Our extensive experiments demonstrate that the proposed method achieves competitive long-term forecasting performance on the benchmark datasets. The detailed ablation studies also show the effectiveness of each module in our model and the generalization on the body pose forecasting task.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Nos. 61906195, 61906193). We would like to thank Xu Han (NYU) for his support.

References

[1] Mikel Ariz, José J. Bengoechea, Arantxa Villanueva, and Rafael Cabeza. A novel 2D/3D database with automatic face annotation for head tracking and pose estimation. *CVIU*, 148:201–210, 2016.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. In *ICCV*, 2019.

[3] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d and 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[6] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. E. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, pages 1–1, 2019.

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[10] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep clustering by gaussian mixture variational autoencoders with graph embedding. In *ICCV*, 2019.

[11] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *IJCV*, 101(3):437–458, 2013.

[12] Farshad Farshadmanesh, Patrick Byrne, Gerald P Keith, Hongying Wang, Brian D Corneil, and J Douglas Crawford. Cross-validated models of the relationships between neck muscle electromyography and three-dimensional head kinematics during gaze behavior. *Journal of neurophysiology*, 107(2):573–590, 2012.

[13] Katerina Fragkiadaki, Sergey Levine, and Jitendra Malik. Recurrent network models for kinematic tracking. In *ICCV*, 2015.

[14] Hao Fu, C. Li, Xiaodong Liu, Jianfeng Gao, A. Çelikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *NAACL-HLT*, 2019.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[16] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *CVPR*, 2017.

[17] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[19] Daniel Hsu. Time series forecasting based on augmented long short-term memory. *arXiv preprint arXiv:1707.00666*, 2017.

[20] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *CVPR*, 2016.

[21] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. Self-supervised intrinsic image decomposition. In *NeurIPS*, 2017.

[22] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[26] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018.

[27] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *CVPR Workshops*, 2017.

[28] Amit Kumar, Azadeh Alavi, and Rama Chellappa. KEPLER: keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *FG*, 2017.

[29] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *CVPR*, 2019.

[30] Stéphane Lathuilière, Rémi Juge, Pablo Mesejo, Rafael Munoz-Salinas, and Radu Horaud. Deep mixture of linear inverse regressions applied to head-pose estimation. In *CVPR*, 2017.

[31] Peipei Li, Xiang Wu, Yibo Hu, Ran He, and Zhenan Sun. M2FPA: A multi-yaw multi-pitch high-quality dataset and benchmark for facial pose analysis. In *ICCV*, 2019.

[32] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander Gaunt. Constrained graph variational autoencoders for molecule design. In *NeurIPS*, 2018.

[33] M. Martin and R. Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *International Conference on 3D Vision*, 2014.

[34] Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017.

[35] Meaghan K. McCluskey and Kathleen E. Cullen. Eye, head, and body coordination during large gaze shifts in rhesus monkeys: Movement kinematics and the influence of posture. *Journal of Neurophysiology*, 97(4):2976–2991, 2007.

[36] S. S. Mukherjee and N. M. Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE TMM*, 17(11):2094–2107, 2015.

[37] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *ECCV*, 2018.

[38] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018.

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[40] Aleksis Pirinen, Erik Gärtner, and Cristian Sminchisescu. Domes to Drones: Self-supervised active triangulation for 3d human pose reconstruction. In *NeurIPS*, 2019.

[41] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPR Workshops*, 2018.

[42] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *NeurIPS*, 2019.

[43] Anke Schwarz, Monica Haurilet, Manuel Martinez, and Rainer Stiefelhagen. Driveahead-a large-scale driver head pose dataset. In *CVPR Workshops*, 2017.

[44] Yuge Shi, Siddharth N, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS*, 2019.

[45] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[46] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, 2015.

[47] Yongbin Sun, Yue Wang, Ziwei Liu, Joshua Siegel, and Sanjay Sarma. PointGrow: Autoregressively learned point cloud generation with self-attention. In *WACV*, 2020.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[49] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, 2016.

[50] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *ICCV*, 2019.

[51] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, and Yadong Mu. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, April 2020.

[52] Keze Wang, Liang Lin, Chenhan Jiang, Chen Qian, and Pengxu Wei. 3D human pose machines with self-supervised learning. *IEEE TPAMI*, 42(5):1069–1082, 2019.

[53] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NeurIPS*, 2017.

[54] Yujia Wang, Wei Liang, Jianbing Shen, Yunde Jia, and Lap-Fai Yu. A deep coarse-to-fine network for head pose estimation from synthetic data. *PR*, 94:196–206, 2019.

[55] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE TNNLS*, 2020.

[56] T. Xue, J. Wu, K. L. Bouman, and W. T. Freeman. Visual dynamics: Stochastic future generation via layered cross convolutional networks. *IEEE TPAMI*, 41(9):2236–2250, 2019.

[57] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018.

[58] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. FSA-Net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *CVPR*, 2019.

[59] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. SSR-Net: A compact soft stagewise regression network for age estimation. In *IJCAI*, 2018.

[60] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, 2019.

[61] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[62] Cagatay Yildiz, Markus Heinonen, and Harri Lahdesmaki. Ode2vae: Deep generative second order odes with bayesian neural networks. In *NeurIPS*, 2019.

[63] Ye Yuan and Kris Kitani. DLow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, 2020.

[64] Xiaohang Zhan, Xingang Pan, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised learning via conditional motion propagation. In *CVPR*, 2019.

[65] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019.

[66] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE TPAMI*, 41(1):162–175, 2019.

[67] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020.