Towards Understanding Future: Consistency Guided Probabilistic Modeling for Action Anticipation

Zhao Xie[1], Yadong Shi[1], Kewei Wu[1]*, Yaru Cheng[1], Dan Guo[1,2,3]*

[1] School of Computer and Information, the Hefei University of Technology, Hefei 230009, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3] Anhui Zhonghuitong Technology Co., Ltd
xiezhao@hfut.edu.cn, hfut.shiyadong@gmail.com, wukewei@hfut.edu.cn, hfut.chengyaru@gmail.com, guodan@hfut.edu.cn

Abstract

Action anticipation aims to infer the action in the unobserved segment (future segment) with the observed segment (past segment). Existing methods focus on learning key past semantics to predict the future, but they do not model the temporal continuity between the past and the future. However, past actions are always highly uncertain in anticipating the unobserved future. The absence of temporal continuity smoothing in the video's past-and-future segments may result in an inconsistent anticipation of future action. In this work, we aim to smooth the global semantics changes in the past and future segments. We propose a Consistency-guided Probabilistic Model (CPM), which focuses on learning the globally temporal probabilistic consistency to inhibit the unexpected temporal consistency. The CPM is deployed on the Transformer architecture, which includes three modules of future semantics estimation, global semantics estimation, and global distribution estimation involving the learning of past-to-future semantics, past-and-future semantics, and semantically probabilistic distributions. To achieve the smoothness of temporal continuity, we follow the principle of variational analysis and describe two probabilistic distributions, i.e., a past-aware distribution and a global-aware distribution, which help to estimate the evidence lower bound of future anticipation. In this study, we maximize the evidence lower bound of future semantics by reducing the distribution distance between the above two distributions for model optimization. Extensive experiments demonstrate that the CPM achieves state-of-the-art performance on Epic-Kitchen100, Epic-Kitchen55, and EGTEA-GAZE.

Introduction

Action anticipation aims to infer the action in the unobserved segment (future segment), which happens after the observed segment (past segment). Action anticipation is an important task in computer vision applications, such as human-robot collaboration (Dessalene et al. 2021), assistive robotics (Liu et al. 2020), smart houses (Damen et al. 2022, 2018), and autonomous vehicle (Zhang et al. 2022). In human activities, action semantics in multiple consecutive segments usually satisfy temporal consistency, i.e., having smooth action variations along the timeline. In order to avoid anticipating undesired actions, action prediction models need to capture the temporal consistency of the global segment (including all the past-future segments). Existing methods learn the temporal semantics with RNN (Furnari and Farinella 2019; Guo et al. 2018; Wang et al. 2018), which uses the memory mechanism to select key semantics in the past segment. Some methods learn the temporal semantics with context model (Guo et al. 2019; Li, Guo, and Wang 2021; Guo, Wang, and Wang 2022; Song et al. 2023a), and Transformer (Girdhar and Grauman 2021; Xu, Li, and Lu 2022; Song et al. 2023b), which enhances the semantics by learning the temporal relation between frames in the global segment. The above methods learn relations between segments, and neglect temporal consistency of the semantics changes from the past to the future. Action anticipation may predict unexpected actions under the temporal inconsistent semantics. Therefore, it is a challenging task to smooth the temporal consistency in the global segment for action anticipation.

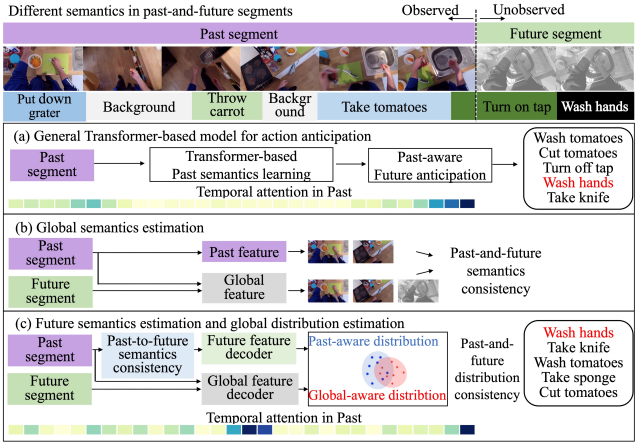

Figure 1: Illustration of action anticipation task and past-and-future temporal consistency. (a) General Transformer-based model for action anticipation. (b) Global semantics estimation. (c) Future semantics estimation and global distribution estimation.

Figure 1 shows that action anticipation can be influenced by different semantics in the global (past-and-future) seg-

*Corresponding author: Kewei Wu, Dan Guo

ments. Unobserved future frames make it easy to take unexpected semantics changes at the boundary between past and future segments, which hinders correct action anticipation. (a) We can use a Transformer-based model to predict future action, but its backbone network pays more attention to high-frequency actions. Since "taking tomatoes" occurs in multiple frames of the past segment and is closely tied to the future segment, the Transformer model pays more temporal attention to "taking tomatoes" and "turning on the tap", which leads to the unexpected future anticipation–"washing tomatoes". (b) To keep the temporal continuity in the video, we consider the global semantics estimation (See Sec.3.2), which is used to constrain the semantics consistency between the past segment and the global (past-and-future) segment. (c) Considering the temporal continuity at the past-to-future boundary, we consider both future semantics estimation module (See Sec.3.1) and global distribution estimation (See Sec.3.3). More importantly, we introduce the probabilistic model into the global distribution estimation. Concretely, in the future semantics estimation, we introduce the probabilistic consistent semantics from past features and use it to learn unknown future features. The future features are anticipated frame-by-frame with multiple consistency-aware decoders. After that, in the global distribution estimation, we introduce a probabilistic global distribution consistency by describing the global semantics with the latent probabilistic variable. Following the variational analysis (Kingma and Welling 2014), we learn the probabilistic distribution consistency with two probabilistic distributions, including a past-aware distribution and a global-aware distribution, which help to estimate the evidence lower bound of the past-aware future anticipation. The past-aware distribution is estimated from past features. The global-aware distribution is estimated with a global feature decoder. We reduce the distribution distance between the above two distributions to maximize the evidence lower bound, which can alleviate the unexpected semantics changes.

In this work, we propose a consistency-guided probabilistic model that focuses on learning probabilistic temporal consistency in the videos. **First**, we design a future semantics estimation module (See Sec.3.1), which considers probabilistic modeling of the past segment to learn past-to-future distribution and finally outputs the anticipated future semantics. The probabilistic distribution can describe the latent feature variation to learn uncertainty-aware future semantics. **Secondly**, we design a global semantics estimation module (See Sec.3.2) that measures the semantics consistency between the past segment and the past-and-future segment. A new loss, global semantics loss L_{sem}, is introduced to describe the constraint of the global semantics changes. **Thirdly**, we design a global distribution estimation module (See Sec.3.3), which introduces a probabilistic global semantics z to approximately maximize past-aware future anticipation as shown in **Figure 2**. Without the probabilistic global semantics z, the direct anticipation needs to directly sample global distribution from past semantics, which is hard to capture the unobserved future clues. Following variational analysis (Kingma and Welling 2014), we approximate the sampling distribution by introducing

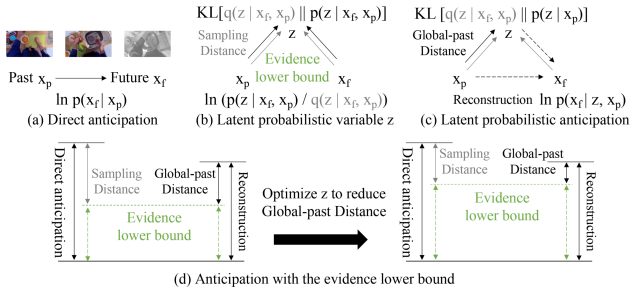

Figure 2: Probabilistic modeling for action anticipation. (a) Direct anticipation. (b) Latent probabilistic variable z. (c) Latent probabilistic anticipation. (d) Anticipation with the evidence lower bound.

global-aware distribution, which can be evidently estimated with probabilistic global semantics. The direct anticipation $lnp(x_f|x_p)$ in Figure 2(a) can be decomposed into a latent estimation $ln(p(z|x_f, x_p)/q(z|x_f, x_p))$ and a sampling distance measurement $KL[q(z|x_f, x_p)||p(z|x_f, x_p)]$ in Figure 2(b). When the sampling of anticipated future features (past-aware) meets the global-aware distribution, the sampling distance gets its smallest value at 0, and the direct anticipation gets its evidence lower bound (ELBO) at the latent estimation. Based on Bayesian theory, the evidence lower bound is a difference term by subtracting global-past distance term $KL[q(z|x_f, x_p)||p(z|x_p)]$ from reconstruction term $lnp(x_f|z, x_p)$ in Figure 2(c). When the global-past distance gets 0, the ELBO gets its maximum value at reconstruction estimation. Turning backward to our framework, we learn the past-aware distribution from past features. and learn the global-aware distribution through a global feature decoder. We use the probabilistic semantic concept to reduce the global-past distance, which can maximize the ELBO. The maximized ELBO can alleviate the neglected effect of unobserved future sampling distribution and increase the probability of correct future anticipation. We design the loss objectives (i.e., a future reconstruction loss L_{rec}, and a distribution distance loss L_{dist}) to optimize the model for smoothing the temporal distribution consistency.

Our contributions are summarized as follows: (1) We propose a consistency-guided probabilistic model, which smooths the global (past-and-future) temporal consistency for action anticipation. We design a future semantics estimation module to learn probabilistic past-to-future consistency, which can alleviate the uncertain past-to-future semantics changes, and ease the unexpected future semantics. (2) We design a global semantics estimation module to constrain the global semantics consistency, which can smooth the temporal inconsistency during past-and-future semantics learning. (3) We design the global distribution estimation module to constrain the probabilistic distribution consistency by reducing the distance between past-aware distribution and global-aware distribution. The global distribution consistency can enhance the future anticipation certainty among the global semantics of past and future segments.

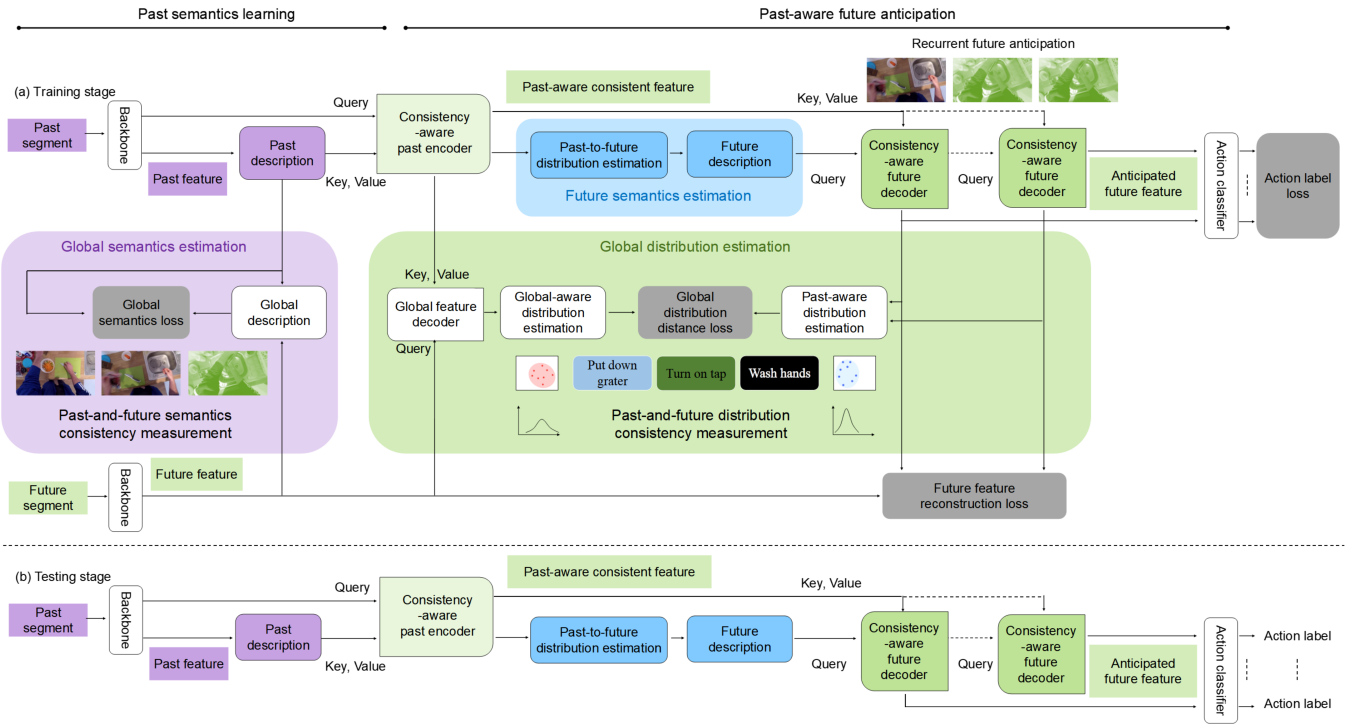

Figure 3: The overview of Consistency-guided Probabilistic Model (CPM). The CPM is deployed on Transformer by introducing three modules, including future semantics estimation, global semantics estimation, and global distribution estimation. In the training stage, the CPM is optimized by integrating three modules for smoothing temporal consistency. In the testing stage, the optimized CPM only uses the future semantics estimation module for action anticipation. The future semantics estimation module anticipates future features by considering past-to-future consistency. The global semantics/distribution estimation modules measure the consistency of past and global (past-and-future) semantics/distribution.

Related Work

Action Anticipation. Action anticipation focuses on learning features in the past segment to predict future action. Some methods learn the temporal feature with temporal convolution network (Carreira and Zisserman 2017). The temporal features can be learned with recurrent model (Liu and Lam 2022; Gao, Yang, and Nevatia 2017; Wu et al. 2021), which can describe the temporal changes in the past segment. The temporal feature can be enhanced with hand movement (Liu et al. 2020), segment-aware feature (Dessalene et al. 2021). The temporal feature can be enhanced by considering UnRolling model (Furnari and Farinella 2019), and high-order frames (Tai et al. 2021). The temporal feature can be learned with latent goal learning (Roy and Fernando 2022a), and semantics-aware contrastive learning (Qi et al. 2023; Zhou, Guo, and Wang 2023), and transferring knowledge (Sener, Saraf, and Yao 2023).

Some methods learn features with temporal Transformer model (Girdhar and Grauman 2021), which can describe the relation between frames as the temporal self-attention. The Transformer can be enhanced with dynamic temporal masks (Xu, Li, and Lu 2022), hierarchical Transformer (Wu et al. 2022), inductive self-attention (Tai et al. 2022), temporal cross-attention (Gong et al. 2022), interactive cross-attention (Roy, Rajendiran, and Fernando 2022), and multiple-scale temporal banks (Sener, Singhania, and Yao 2020). The Transformer can be enhanced with cross-modality self-attention (Gu et al. 2021; Zhong et al. 2023). The above Transformers neglect to smooth the unexpected temporal consistency, which hinders explaining the unexpected action anticipation.

Probabilistic Semantics Modeling. The probabilistic model can describe the feature variation as the distribution of semantics. The variant features can predict the multiple possible labels (Yang et al. 2023). The probabilistic prediction has been described in the Bayesian language model (Xue et al. 2022; Zhang et al. 2021) and the complex action recognition model (Guo, Wang, and Ji 2022). Some methods learn probabilistic prediction with the latent variables (Zheng et al. 2022; Itkina et al. 2020; Pambala, Dutta, and Biswas 2020; Zhang et al. 2020). Abstract Goal (Roy and Fernando 2022b) learns temporal features by introducing a probabilistic recurrent network. Abstract Goal learns the Gaussian distribution of latent semantics, and considers the distribution loss between the future semantics and the past semantics. Unlike the Abstract Goal, our work introduces a probabilistic recurrent Transformer, which focuses on smoothing the unexpected past-to-future consistency, and smoothing temporal consistency in semantics and distributions.

Method

As shown in **Figure 3**, we introduce the procedure of the consistency-guided probabilistic model (CPM), which contains a training stage, and a testing stage. In the training stage, we adopt the Transformer (Vaswani et al. 2017) to construct a future semantics estimation module, a global semantics estimation module, and a global distribution estimation module. In the testing stage, the model only uses the future semantics estimation module. For feature extraction, we take the temporal shift model (Lin, Gan, and Han 2019) as the backbone to extract the past segment feature $f_p \in R^{T_p \times D}$, and the future segment feature $f_f \in R^{T_f \times D}$, where T_p, T_f denote the frame number of each segment. D denotes the feature dimension.

Future Semantics Estimation

The future semantics estimation module is designed to learn the probabilistic past-to-future semantics. We use the probabilistic past-to-future semantics to anticipate the future features frame-by-frame with a consistency-aware future decoder. The decoder considers the probabilistic past-to-future semantics to smooth the unexpected semantics changes, which can also smoothly anticipate the semantics in the future segment.

At first, we use a consistency-aware past encoder to learn the temporal relation between past features. Following Vision Transformer (Dosovitskiy et al. 2021), we append a learnable class token f_{token} before past feature. Given the past feature f_p, an Fully Connected (FC) layer is used to learn the channel relations in one frame, and projects the past feature into $f'_p \in R^{T_p \times D}$. The Query takes the original past feature $f_{cl,p} = [f_{token}, f_p]$. The Key and Value take the projected feature $f'_{cl,p} = [f_{token}, f'_p]$. In the Transformer encoder layer, the temporal feature is estimated as:

$$\begin{cases} Q_{en} = f_{cl,p}W_Q, K_{en} = f'_{cl,p}W_K, V_{en} = f'_{cl,p}W_V, \\ Q'_{en} = softmax(\frac{Q_{en} \cdot K_{en}^T}{\sqrt{D}})V_{en} + Q_{en} \\ Q''_{en} = FFN(Q'_{en}) \end{cases} \tag{1}$$

where W_Q, W_K, W_V are learnable parameters. FFN is the feed-forward network. The next encoder layer uses the same Key and Value in the first layer, and uses the Query with the output of the previous layer. The regressive Transformer encoder finally outputs the encoded past feature $x_p \in R^{T_p \times D}$.

The future semantics estimation module is designed to describe the past-to-future consistency with Gaussian distribution. At the last frame T_p in the past segment, we estimate the Gaussian distribution by considering the frame feature and the class token. The class token is the first feature in the encoded past feature $x_{token} = x_{p,0}$. The *mean* values are estimated with multiple layer perceptron (MLP) as $\mu_{T_p} = MLP(x_{p,T_p} + x_{token}) \in R^{1 \times D}$. The *standard* values are estimated as $\sigma_{T_p} = MLP(x_{p,T_p} + x_{token}) \in R^{1 \times D}$. We introduce a Gaussian variable to describe the unexpected consistency $\epsilon_{T_p} \sim Gaussian(\mathbf{0}, \mathbf{I})$. The past-aware feature at timestamp T_p is described as:

$$x'_{p,T_p} = \sigma_{T_p} \cdot \epsilon_{T_p} + \mu_{T_p} \tag{2}$$

After obtaining x'_{p,T_p}, we further use a consistency-aware future decoder to estimate the next future feature. We stack multiple decoders recurrently to learn future features frame-by-frame. Following Transformer (Vaswani et al. 2017), the decoder layer uses the feature directly without the class token. The Query takes the past-aware feature x'_{p,T_p}. The Key and Value take the past feature x_p. In the first decoder layer, the temporal feature is estimated as:

$$\begin{cases} Q_{re} = x'_{p,T_p}W_Q, K_{re} = x_pW_K, V_{re} = x_pW_V, \\ Q'_{re} = softmax(\frac{Q_{re} \cdot K_{re}^T}{\sqrt{D}})V_{re} + Q_{re} \\ Q''_{re} = FFN(Q'_{re}) \end{cases} \tag{3}$$

where W_Q, W_K, W_V are learnable parameters. FFN is the feed-forward network. The next future decoder layer uses the same Key and Value in the first layer, and uses the Qurey with the output of the previous layer. The first unit outputs the feature of the T_p+1 frame. The Multiple Transformer decoders estimate the future features from x'_{p,T_p+1} to x'_{p,T_p+T_f} frame-by-frame.

Global Semantics Estimation

The global semantics estimation module is introduced to learn the global (past-and-future) semantics consistency, which is estimated with the distance between past semantics and global semantics. We consider the distance as the global semantics loss. We reduce global semantics loss to keep the global semantics consistency. The global semantics consistency can smooth the past feature, which helps to learn the evidence lower bound of the future feature anticipation.

To describe the global feature of the input video, we concatenate the past feature f_p with the future feature f_f as the global feature: $f_g = concat(f_p, f_f)$. We trim the global feature from the last frames to align the temporal length of the past feature. We average the past feature and the aligned feature, and project the feature with an FC layer to describe the global feature $f'_g = FC((f_p + f_{g,T_f+1:T_f+T_p})/2) \in R^{T_p \times D}$. We reduce the distance between the projected global feature and the projected past feature to learn the global semantics consistency. The global semantics consistency can use the global feature to align the past feature. We estimate their distance by using the negative value of the Jaccard vector similarity, and use the distance as the global semantics loss:

$$L_{sem} = exp(-\sum_{t=1}^{T_p} \frac{2 \cdot f'_{p,t} \cdot (f'_{g,t})^T}{f'_{p,t} \cdot (f'_{p,t})^T + f'_{g,t} \cdot (f'_{g,t})^T}) \tag{4}$$

Global Distribution Estimation

Unlike the above global semantics estimation, this section estimates *global distributions* to measure the temporal consistency. The global distribution estimation module learns the probabilistic global (past-and-future) distribution consistency with probabilistic global semantics. Without the probabilistic global semantics, the direct past-aware future anticipation needs the sampling distribution of global semantics, which is hard to estimate due to unobserved future frames. Following the variational analysis (Kingma and Welling

2014) as discussed in **Introduction**, we introduce the probabilistic global semantics. The probabilistic global semantics provides the evidence to estimate the global-aware distribution, which can be used as the approximate estimation of the unobserved sampling distribution. The probabilistic global semantics helps to decompose the direct anticipation into the latent anticipation estimation and the sampling distance estimation. When the sampling distribution is the same as the global-aware distribution, the sampling distance gets its smallest value at 0, and the direct anticipation gets its evidence lower bound at the latent anticipation estimation. We maximize the evidence lower bound, which can enhance the past-aware future anticipation.

Given the future feature x_f and the past features x_p, the direct anticipation is $lnp(x_f|x_p)$. We address the evidence lower bound (ELBO) of direct anticipation in the supplementary materials. We introduce the probabilistic global semantics as z. Given the sampling distribution $q(z|x_f, x_p)$, and the global-aware distribution $p(z|x_f, x_p)$, the evidence lower bound is $ln(p(z|x_f, x_p)/q(z|x_f, x_p))$. Based on the Bayesian theory (Kingma and Welling 2014), the evidence lower bound is the difference term by subtracting the distribution distance term from the reconstruction term:

$$\begin{aligned} & lnp(x_f|x_p) = \mathbb{E}_{q(z|x_f,x_p)}[lnp(x_f|x_p)] \\ & = \mathbb{E}_{q(z|x_f,x_p)}[ln\tfrac{p(x_f,z|x_p)}{q(z|x_f,x_p)}] \\ & \quad + \underbrace{\mathbb{D}_{KL}[q(z|x_f,x_p)||p(z|x_f,x_p)]}_{\geq 0} \\ & \geq \mathbb{E}_{q(z|x_f,x_p)}[ln\tfrac{p(x_f,z|x_p)}{q(z|x_f,x_p)}] := ELBO \\ & = \underbrace{\mathbb{E}_{q(z|x_f,x_p)}[lnp(x_f|z,x_p)]}_{Reconstruction} - \underbrace{\mathbb{D}_{KL}[q(z|x_f,x_p)||p(z|x_p)]}_{Distribution\ distance} \end{aligned} \tag{5}$$

where the $\mathbb{D}_{KL}[q(\cdot)||p(\cdot)]$ is the Kullback-Leibler divergence to describe the distance between two probabilities.

Future Reconstruction Loss. The future reconstruction loss is used to estimate the reconstruction term in ELBO. We maximize the reconstruction term by reducing the distance between the anticipated future feature $x'_{p,T_p+1:T_p+T_f}$ and the feature of the future segment $f_{f,1:T_f}$. We use Jaccard vector similarity to measure the reconstruction loss:

$$L_{rec} = exp(-\sum_{t=1}^{T_f} \frac{2 \cdot x'_{p,T_p+t} \cdot (f_{f,t})^T}{x'_{p,T_p+t} \cdot (x'_{p,T_p+t})^T + f_{f,t} \cdot (f_{f,t})^T}) \tag{6}$$

Distribution Distance Loss. The distribution distance loss is used to estimate the distribution distance term in ELBO, which is between the past-aware distribution and the global-aware distribution. We estimate the Gaussian-based past-aware distribution with future features anticipated from the past features. In frame $t \in [T_p + 1, T_p + T_f]$, the *mean* values are estimated with multiple layer perceptron as $\mu_{p,t} = MLP(x'_{p,t}) \in R^{1\times D}$. The *standard* value are estimated as $\sigma_{p,t} = MLP(x'_{p,t}) \in R^{1\times D}$. The probabilistic global semantics is learned as the probability following the past-aware distribution: $P(z|x'_{p,t}) \sim Gaussian(\mu_{p,t}, \sigma^2_{p,t})$.

To estimate the global-aware distribution, we introduce a global feature decoder. Following Transformer (Vaswani et al. 2017), the decoder uses the feature directly without considering the class token. The Query takes the future feature f_f. The Key and Value take the past feature x_p. The Transformer decoder contains multiple regressive decoder layers too. In the first layer of the Transformer decoder, the temporal feature is estimated as:

$$\begin{cases} Q_g = f_f W_Q, K_g = x_p W_K, V_g = x_p W_V, \\ Q'_g = softmax(\frac{Q_g \cdot K_g^T}{\sqrt{D}}) V_g + Q_g \\ Q''_g = FFN(Q'_g) \end{cases} \tag{7}$$

where W_Q, W_K, W_V are learnable parameters. FFN is the feed-forward network. The next decoder layer uses the same Key and Value in the first layer, and uses the Query with the output of the previous layer. The Transformer decoder uses multiple layers to output the final global feature $x'_g \in R^{T_f \times D}$.

We estimate the global-aware distribution with Gaussian distribution. In frame $t \in [T_p + 1, T_p + T_f]$, the *mean* values are estimated as $\mu_{g,t} = MLP(x'_{g,t})$. The *standard* value are estimated as $\sigma_{g,t} = MLP(x'_{g,t})$. The probabilistic global semantics depending on global features follows the global-aware distribution as: $P(z|x'_{g,t}) \sim Gaussian(\mu_{g,t}, \sigma^2_{g,t})$.

We address the Kullback-Leibler divergence between two Gaussian-based distributions in the supplementary materials. Given the estimated past-aware distribution and global-aware distribution, the global distribution distance loss is the summation of Kullback-Leibler divergence at each timestamp t and each dimension d as:

$$L_{\text{dist}} = \sum_{t=T_p+1}^{T_p+T_f} \sum_{d=1}^{D} \left[\ln \frac{\sigma_{p,t,d}}{\sigma_{g,t,d}} - \frac{1}{2} + \frac{\sigma^2_{g,t,d} + (\mu_{g,t,d} - \mu_{p,t,d})^2}{2\sigma^2_{p,t,d}} \right] \tag{8}$$

Model Optimization

Our model is optimized by considering both the temporal consistency loss and the action prediction loss. The temporal consistency loss contains a global semantics loss L_{sem} (See Sec. 3.2), a future reconstruction loss L_{rec} (See Sec. 3.3), and a distribution distance loss L_{dist} (See Sec. 3.3). The action prediction loss is the base term for this task. To be specific, the action classifier uses two fully connected layers to predict the future action. At timestamp t in the anticipation segment, the predicted action labels include the noun label $y_{noun,t}$, the verb label $y_{verb,t}$, and the action label $y_{act,t}$. Given the ground truth of the noun label $y^{gt}_{noun,t}$, the verb label $y^{gt}_{verb,t}$, and the action label $y^{gt}_{act,t}$, we use the cross-entropy loss to estimate the difference between the predicted and the ground truth labels. For each video, the noun label loss is $L_{noun} = \sum_t -y^{gt}_{noun,t} \cdot log(y_{noun,t})$; the verb label loss is $L_{verb} = \sum_t -y^{gt}_{verb,t} \cdot log(y_{verb,t})$; the action label loss is $L_{act} = \sum_t -y^{gt}_{act,t} \cdot log(y_{act,t})$. The above losses compose the action prediction loss as $L_{base} = L_{noun} + L_{verb} + L_{act}$.

For model optimization, the total optimization objective is formulated as follows:

$$L_{total} = L_{base} + L_{sem} + \underbrace{L_{rec} + L_{dist}}_{ELBO}. \tag{9}$$

Dataset	Segments	Classes	τ_a	Metric
EK100	90.0K	3,807	1.0s	Recall@5
EK55	39.6K	2,513	1.0s	Top-1/5
EG+	10.3K	106	0.5s	Top-1, CM Top-1

Table 1: Datasets used for action anticipation.

Method	EK 100	EK 55	
	Recall@5	Top 1	Top 5
Modality-RGB			
ED	–	25.8	–
HORST	13.2	–	–
HORST-url	13.2	–	–
DCR-1s (TSN)	14.6	13.6	30.8
AVT (AVT-b)	14.9	–	–
MeMViT-16	15.1	–	–
DCR-1s (TSM)	16.1	16.1	33.1
Ours	**17.2**	**17.2**	**35.2**
Modality-RGB+OBJ+Flow			
RULSTM	14.0	15.3	35.3
ActionBanks	14.7	15.1	35.6
HRO	–	–	37.4
AVT+	14.8	16.6	37.6
AVT++	15.9	–	–
TransAction	16.6	–	–
Ego-OMG	–	19.2	–
DCR-1s (TSN)	18.3	19.2	41.2
Ours	**19.4**	**20.1**	**43.6**

Table 2: Comparison with state-of-the-art methods on the validation set of EK100 (Class-mean recall@5 of action prediction) and the validation set of EK55 (Top-1 accuracy and Top-5 accuracy of action prediction.)

Experiments

Datasets and Implementation Details

Datasets. **Table 1** shows three datasets and their evaluation metrics. τ_a represents the interval between past segment and future action. **EpicKitchens-100** (EK100) (Damen et al. 2022) is the egocentric (first-person) video dataset of cooking activities. **EpicKitchens-55** (EK55) (Damen et al. 2018) is an earlier version of EK100. For EK100/EK55, we report performance with the standard train, validation, and test splits from (Damen et al. 2022) and (Furnari and Farinella 2019) respectively. **EGTEA Gaze+** (EG+) (Li, Liu, and Rehg 2018) is another first-person anticipation dataset. Following (Liu et al. 2020), we use the split 1 (Li, Liu, and Rehg 2018) of the dataset.

Metrics. The anticipation metrics are estimated with action prediction, including Top-1/5, class mean (CM) Top-1, and Recall@5. The action prediction is decomposed into verb prediction and noun prediction (Girdhar and Grauman 2021). The smoothness metrics are estimated with the temporal consistency curve, which is learned by averaging the features across channels, Specifically, the mean absolute temporal difference (MATD) is the mean value of the absolute temporal difference between neighbor frames. The mean curvature (MC) is the mean value of the curvature of the curve.

Method	Top-1	CM Top-1
Modality-RGB+Flow		
I3D-Res50	34.8	23.2
FHOI	36.6	25.3
RULSTM	38.6	–
AVT-h	39.8	28.3
AVT	43.0	35.2
Abstract Goal	49.8	37.4
Ours	**50.3**	**39.1**

Table 3: Comparison with state-of-the-art methods on EG+. Top-1 accuracy and Class Mean (CM) Top-1 accuracy of action prediction.

Encoder	Future	Decoder	Verb	Noun	Act.
3 layers	✗	3 layers	27.2	25.5	13.1
3 layers	✓	3 layers	34.5	34.0	16.4
3 layers	✓	4 layers	**36.9**	**36.2**	**17.2**
4 layers	✓	3 layers	36.1	35.5	16.8
4 layers	✓	4 layers	33.1	33.0	15.6

Table 4: The effect of the future semantics estimation.

Implementation Details. Following (Xu, Li, and Lu 2022), we use TSM as the backbone (Lin, Gan, and Han 2019). For EK100/EK55, the past and future segments have 30 frames and 8 frames, respectively. For EG+, the past and future segments have 20 frames and 8 frames, respectively. We use the Transformer encoder with a 3-layer, 16-head, 1024-dimensional model, and the Transformer decoder with a 4-layer, 16-head, 1024-dimensional model optimized by AdamW (Loshchilov and Hutter 2019). We initialize the Transformer from scratch (Xu, Li, and Lu 2022). For EK100/EK55, we set the learning rate to 1e-4, batch size to 128, and training epoch to 100. For EG+, we set the learning rate to 5e-5, batch size to 512, and training epoch to 50.

Comparison with State-of-the-Art

Table 2 shows the performance on **EK100** and **EK55**. The first part methods are Transformer-based models with RGB features. DCR-1s (Xu, Li, and Lu 2022) learns features with multiple temporal masks. Our method smooths global temporal consistency from past semantics learning to future semantics learning, and outperforms the above Transformer-based methods. The second part methods use the RGB, object, and Flow features. Our model optimizes the probabilistic global semantics to maximize the future anticipation, and outperforms the model with distribution consistency. **Table 3** shows the comparision on **EG+**. Our method learns smoothed global temporal consistency, and outperforms AVT and Abstract Goal.

Ablation Study

To verify the probabilistic temporal consistency learning, we perform ablation studies using RGB features learned with TSM backbone (Lin, Gan, and Han 2019) on EK100.

The Effect of Future Semantics Estimation. **Table 4** shows the Top-1 accuracy of the verb label (Verb), the noun label (Noun), and the action label (Act.) on overall classes.

L_{sem}	L_{rec}	Verb	Noun	Act.
Negative dot	Negative dot	33.1	33.4	15.3
L2 distance	L2 distance	34.0	34.1	16.0
Negative cosine	Negative cosine	34.3	34.3	16.2
Negative Jaccard	Negative Jaccard	**36.9**	**36.2**	**17.2**

Table 5: The effect of the distance estimation in losses.

L_{base}	L_{sem}	L_{dist}	L_{rec}	Verb	Noun	Act.
✓	–	–	–	32.1	32.4	14.6
✓	✓	–	–	33.5	33.5	15.4
✓	✓	✓	–	36.2	35.7	16.8
✓	✓	✓	✓	**36.9**	**36.2**	**17.2**

Table 6: The effect of the model losses.

When we add the probabilistic past-to-future semantics to anticipate the future features, the model increases the performance compared with the model without probabilistic past-to-future semantics. The model with a 3-layer decoder and a 4-layer encoder gets the best performance.

The Effect of Distance Estimation in Feature Losses. **Table 5** shows the performance with different distance estimations in global semantics loss L_{sem} and reconstruction loss L_{rec}. To estimate the distance between two feature matrixes, we temporally split them into vectors at each timestamp. In the Jaccard function, the denominators can describe the number of semantics in two feature vectors. The model with the negative Jaccard function in two losses gets the best performance.

The Effect of Model Losses. **Table 6** shows the performances with different losses. The base losses can reduce the error of the verb/noun/action label predictions. The semantics loss can embed global semantics for past semantics learning. The reconstruction loss can embed the future feature into the feature anticipated from the past feature. The distance loss can embed global-aware distribution into past-aware distribution for probabilistic distribution learning. We use the above four losses to get the best performance.

Visualization

Visualization Analysis of Temporal Consistency Learning. **Figure 4** shows two instances of action anticipation. We compare our CPM with the base model, which removes future semantics estimation, global semantics estimation, and global distribution estimation from our CPM. As the past-future features in Figure 4 (left), our global semantics estimation helps to learn smoothed past features, which have small distances between each other. As the feature curve in Figure 4 (middle), we visualize the curve with the semantic probability by averaging the features across channels at each timestamp, which can verify the temporal smoothness of past and anticipated future actions. In both two videos, the base model takes a large change at the past-to-future boundary and has the sharp peak on the future frame. Table 7 shows the mean absolute temporal difference (MATD) and the mean curvature (MC) on EK100. Our CPM can enhance the smoothness of the temporal consistency. As the distribution in Figure 4 (right), we notice the global-aware distribution and the past-aware distribution in the base model take far apart. Our CPM enforces the distribution consistency with the probabilistic model, and the above two distributions are more consistent by overlapping each other. As the attention in Figure 4 (upper), we further show the temporal attention along the timeline of the video. In (a), the base model pays attention to "taking tomatoes" and "turning on tap", and mispredicts the future action as "washing hands". Our model suggests more attention on "throwing carrot" in the trash can, which may dirty hands. To clean the hands, our model predicts the correct future action as "washing hands". In (b), Our model smooths the temporal consistency and keeps attention on the "opening container". The temporal consistency of "opening container" suggests that the future semantics should be highly relevant to "closing container".

Base / CPM	Fig. 4(a)	Fig. 4(b)	EK100
MATD ↓	0.146 / **0.072**	0.128 / **0.075**	0.125 / **0.067**
MC ↓	0.222 / **0.116**	0.218 / **0.108**	0.192 / **0.097**

Table 7: The smoothness metrics of temporal consistency.

Figure 4: Temporal consistency in semantics and distributions of Consistency-guided Probabilistic Model (CPM).

Conclusion

This work proposes a consistency-guided probabilistic model, which learns probabilistic global temporal consistency to smooth the unexpected temporal continuity for action anticipation. We learn probabilistic past-to-future consistency to anticipate smoothed future features. We learn global semantics consistency to embed the global semantics for past semantics learning. We learn probabilistic global distribution consistency to embed the global-aware distribution for past-aware future anticipation.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB4500600), and the National Natural Science Foundation of China (62272144 and U20A20183).

References

Carreira, J.; and Zisserman, A. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 4724–4733.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. *CoRR*, abs/1804.02748.

Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.*, 130(1): 33–55.

Dessalene, E.; Devaraj, C.; Maynord, M.; Fermüller, C.; and Aloimonos, Y. 2021. Forecasting Action through Contact Representations from First Person Video. *CoRR*, abs/2102.00649.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Furnari, A.; and Farinella, G. M. 2019. What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6251–6260.

Gao, J.; Yang, Z.; and Nevatia, R. 2017. RED: Reinforced Encoder-Decoder Networks for Action Anticipation. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*.

Girdhar, R.; and Grauman, K. 2021. Anticipative Video Transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 13485–13495.

Gong, D.; Lee, J.; Kim, M.; Ha, S. J.; and Cho, M. 2022. Future Transformer for Long-term Action Anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 3042–3051.

Gu, X.; Qiu, J.; Guo, Y.; Lo, B.; and Yang, G. 2021. TransAction: ICL-SJTU Submission to EPIC-Kitchens Action Anticipation Challenge 2021. *CoRR*, abs/2107.13259.

Guo, D.; Li, K.; Zha, Z.; and Wang, M. 2019. DADNet: Dilated-Attention-Deformable ConvNet for Crowd Counting. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, 1823–1832.

Guo, D.; Wang, H.; and Wang, M. 2022. Context-Aware Graph Inference With Knowledge Distillation for Visual Dialog. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10): 6056–6073.

Guo, D.; Zhou, W.; Li, H.; and Wang, M. 2018. Hierarchical LSTM for Sign Language Translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 6845–6852.

Guo, H.; Wang, H.; and Ji, Q. 2022. Uncertainty-Guided Probabilistic Transformer for Complex Action Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 20020–20029.

Itkina, M.; Ivanovic, B.; Senanayake, R.; Kochenderfer, M. J.; and Pavone, M. 2020. Evidential Sparsification of Multimodal Latent Spaces in Conditional Variational Autoencoders. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Li, K.; Guo, D.; and Wang, M. 2021. Proposal-Free Video Grounding with Contextual Pyramid Network. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021*, 1902–1910.

Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, 639–655.

Lin, J.; Gan, C.; and Han, S. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 7082–7092.

Liu, M.; Tang, S.; Li, Y.; and Rehg, J. M. 2020. Forecasting Human-Object Interaction: Joint Prediction of Motor Attention and Actions in First Person Video. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, 704–721.

Liu, T.; and Lam, K. 2022. A Hybrid Egocentric Activity Anticipation Framework via Memory-Augmented Recurrent and One-shot Representation Forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 13894–13903.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Pambala, A. K.; Dutta, T.; and Biswas, S. 2020. Generative Model with Semantic Embedding and Integrated Classifier for Generalized Zero-Shot Learning. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, 1226–1235.

Qi, Z.; Wang, S.; Su, C.; Su, L.; Huang, Q.; and Tian, Q. 2023. Self-Regulated Learning for Egocentric Video Activity Anticipation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6): 6715–6730.

Roy, D.; and Fernando, B. 2022a. Action anticipation using latent goal learning. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, 808–816.

Roy, D.; and Fernando, B. 2022b. Predicting the Next Action by Modeling the Abstract Goal. *CoRR*, abs/2209.05044.

Roy, D.; Rajendiran, R.; and Fernando, B. 2022. Interaction Visual Transformer for Egocentric Action Anticipation. *CoRR*, abs/2211.14154.

Sener, F.; Saraf, R.; and Yao, A. 2023. Transferring Knowledge From Text to Video: Zero-Shot Anticipation for Procedural Actions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6): 7836–7852.

Sener, F.; Singhania, D.; and Yao, A. 2020. Temporal Aggregate Representations for Long-Range Video Understanding. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, volume 12361 of *Lecture Notes in Computer Science*, 154–171.

Song, P.; Guo, D.; Cheng, J.; and Wang, M. 2023a. Contextual Attention Network for Emotional Video Captioning. *IEEE Trans. Multim.*, 25: 1858–1867.

Song, P.; Guo, D.; Yang, X.; Tang, S.; Yang, E.; and Wang, M. 2023b. Emotion-Prior Awareness Network for Emotional Video Captioning. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 589–600.

Tai, T.; Fiameni, G.; Lee, C.; and Lanz, O. 2021. Higher Order Recurrent Space-Time Transformer. *CoRR*, abs/2104.08665.

Tai, T.; Fiameni, G.; Lee, C.; See, S.; and Lanz, O. 2022. Inductive Attention for Video Action Anticipation. *CoRR*, abs/2212.08830.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Wang, S.; Guo, D.; Zhou, W.; Zha, Z.; and Wang, M. 2018. Connectionist Temporal Fusion for Sign Language Translation. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, 1483–1491.

Wu, C.; Li, Y.; Mangalam, K.; Fan, H.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. MeMViT: Memory-Augmented Multiscale Vision Transformer for Efficient Long-Term Video Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 13577–13587.

Wu, Y.; Zhu, L.; Wang, X.; Yang, Y.; and Wu, F. 2021. Learning to Anticipate Egocentric Actions by Imagination. *IEEE Trans. Image Process.*, 30: 1143–1152.

Xu, X.; Li, Y.; and Lu, C. 2022. Learning to Anticipate Future with Dynamic Context Removal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 12724–12734.

Xue, B.; Hu, S.; Xu, J.; Geng, M.; Liu, X.; and Meng, H. 2022. Bayesian Neural Network Language Modeling for Speech Recognition. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 2900–2917.

Yang, W.; Zhang, T.; Zhang, Y.; and Wu, F. 2023. Uncertainty Guided Collaborative Training for Weakly Supervised and Unsupervised Temporal Action Localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 5252–5267.

Zhang, J.; Fan, D.; Dai, Y.; Anwar, S.; Saleh, F. S.; Zhang, T.; and Barnes, N. 2020. UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 8579–8588.

Zhang, S.; Abdel-Aty, M. A.; Wu, Y.; and Zheng, O. 2022. Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation. *IEEE Trans. Intell. Transp. Syst.*, 23(3): 2331–2339.

Zhang, S.; Fan, X.; Chen, B.; and Zhou, M. 2021. Bayesian Attention Belief Networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 12413–12426.

Zheng, Y.; He, T.; Qiu, Y.; and Wipf, D. P. 2022. Learning Manifold Dimensions with Conditional Variational Autoencoders. In *NeurIPS*.

Zhong, Z.; Schneider, D.; Voit, M.; Stiefelhagen, R.; and Beyerer, J. 2023. Anticipative Feature Fusion Transformer for Multi-Modal Action Anticipation. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, 6057–6066.

Zhou, J.; Guo, D.; and Wang, M. 2023. Contrastive Positive Sample Propagation Along the Audio-Visual Event Line. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6): 7239–7257.