

# Multimodal Variational Auto-encoder based Audio-Visual Segmentation

Yuxin Mao<sup>1</sup> Jing Zhang<sup>2</sup> Mochu Xiang<sup>1</sup> Yiran Zhong<sup>3</sup> Yuchao Dai<sup>1†</sup>

<sup>1</sup>Northwestern Polytechnical University & Shaanxi Key Laboratory of Information Acquisition and Processing

<sup>2</sup>Australian National University <sup>3</sup>Shanghai AI Laboratory

 <https://github.com/OpenNLPLab/MMVAE-AVS>

 <https://npucvr.github.io/MMVAE-AVS>

## Abstract

We propose an Explicit Conditional Multimodal Variational Auto-Encoder (ECMVAE) for audio-visual segmentation (AVS), aiming to segment sound sources in the video sequence. Existing AVS methods focus on implicit feature fusion strategies, where models are trained to fit the discrete samples in the dataset. With a limited and less diverse dataset, the resulting performance is usually unsatisfactory. In contrast, we address this problem from an effective representation learning perspective, aiming to model the contribution of each modality explicitly. Specifically, we find that audio contains critical category information of the sound producers, and visual data provides candidate sound producer(s). Their shared information corresponds to the target sound producer(s) shown in the visual data. In this case, cross-modal shared representation learning is especially important for AVS. To achieve this, our ECMVAE factorizes the representations of each modality with a modality-shared representation and a modality-specific representation. An orthogonality constraint is applied between the shared and specific representations to maintain the exclusive attribute of the factorized latent code. Further, a mutual information maximization regularizer is introduced to achieve extensive exploration of each modality. Quantitative and qualitative evaluations on the AVSBench demonstrate the effectiveness of our approach, leading to a new state-of-the-art for AVS, with a 3.84 mIOU performance leap on the challenging MS3 subset for multiple sound source segmentation.

## 1. Introduction

Audio-visual data can work collaboratively towards a better perception of the scene. The audio-visual segmentation (AVS) [1, 2] task aims to segment the objects from the video sequence that producing the sound in the audio. On

the one hand, the audio data provides category information for the localization of the object in the video. On the other hand, the visual data provides a sound producer pool with precise structure information of the foreground (sound producer(s)). Different from conventional multimodal settings, where each modality can be used individually for the target task, audio in AVS task serves as “command” to localize and segment the sound producer(s) from the visual data. In this case, the contribution of audio should be extensively explored for accurate foreground segmentation.

The baseline model [1] focuses on implicit feature fusion via audio-visual cross attention. It purely relies on fitting the discrete samples in the dataset. Although reasonable performance is obtained, there are no constraints to guarantee the contribution of each modality, making it hard to decide if the audio data is effectively used, as the model can directly regress the final segmentation maps by only taking the video as input. Fortunately, we discover that each modality for AVS contains both shared and specific information. For example, the audio data includes both the information from the sound producers and the background noise, while the visual data shows the appearance of the entire scene, where the sound producers only take a small portion of it. In our specific task setting, we find modality factorization is suitable to model both the modality-shared representation, *i.e.* information of the sound producers, and the modality-specific representation toward a better understanding of the contribution of each modality.

The straightforward solution to learn the feature representation of the input data is through an auto-encoder (AE) framework. However, AE is mainly used for data compression, as the learned feature space is not continuous, which cannot provide a rich semantic correlation of the data. Differently, with latent space regularization, *e.g.* the latent space is assumed to be Gaussian in variational auto-encoder (VAE) [3], VAE obtains semantic meaningful latent space, which is continuous, and it is also the basic requirement for reliable latent space factorization.

To learn the semantic correlated feature representation of the AVS data, we propose an Explicit Conditional Mul-

<sup>†</sup> Corresponding author (daiyuchao@gmail.com).

This work was done when Yuxin Mao was an intern at Shanghai AI Laboratory.

timodal Variational Auto-Encoder (ECMVAE) for audio-visual segmentation to learn both the *shared* and the *specific* representation in the latent space of each modality. Our model is built upon a multimodal variational auto-encoder [4, 5], with the Jensen-Shannon divergence to achieve a trade-off between sampling efficiency and sample quality. Based on the latent space factorization, we impose constraints for the shared and specific representations to explicitly maximize the contribution of each modality.

Specifically, we first assume that one latent code of the factorized representation should contain independent information compared to others. Furthermore, for the fused representation, we further claim that it should be more informative for the target task compared with each modality. To achieve the former, we propose an information orthogonality constraint between the factorized representations of each modality to ensure that the modality-shared and modality-specific representations capture different aspects of the audio-visual input. For the latter, we fuse the factorized representations of each modality to construct a fused space. Then we introduce a mutual information maximization regularizer between the fused representations of each modality to extensively explore the contribution of each modality. Extensive experimental results demonstrate that our ECMVAE achieves *state-of-the-art* AVS performance. Our pipeline achieves a 3.84 mIOU improvement for the challenging multiple sound source segmentation.

We summarize our main contributions as:

- An explicit semantic correlated feature representation learning framework for audio-visual segmentation is proposed with latent space factorization to capture both the modality-shared and specific representations.
- Based on the latent space factorization, we introduce a unimodal orthogonality constraint between the shared and specific representations and the cross-modal mutual-information maximization regularizer to extensively explore the contribution of each modality.
- State-of-the-art segmentation performance is achieved, showing both the effectiveness of each module and the contribution of each modality.

## 2. Related Work

**Audio-Visual Segmentation.** The Audio-Visual Segmentation (AVS) task is newly proposed, aiming to localize the sound producers with pixel-wise segmentation masks. Zhou *et al.* [1] propose an AVSBench dataset for audio-visual segmentation and provide a simple baseline based on temporal pixel-wise audio-visual interaction (TPAVI), which is a cross-modal attention [6] based fusion strategy. The other audio-visual collaboration tasks can be classified as audio-visual correspondence (AVC) [7, 8], event localization (AVEL) [9–14], event parsing (AVP) [15–17], *etc.* These methods require the fusion of audio and visual sig-

nals. Such as audio-visual similarity modeling by computing the correlation matrix [7, 8], audio-visual cross attention [13, 14, 18], audio-guided Grad-CAM [19], or using a multimodal transformer for modeling the long-range dependencies between elements across modalities directly [20, 21]. However, the challenge and uniqueness of the AVS task are how to map the audio signals to *fine-grained* visual cues, *i.e.* per-pixel segmentation maps. This will rely on reliable modeling of visual and audio signals, as well as more effective fusion strategies.

**Multimodal Variational Auto-encoders.** The Multimodal Variational Auto-encoders (MVAEs) [4, 22–25] are a type of latent variable generative model to learn more generalizable representations from diverse modalities. To achieve this, the core of MVAE is the joint distribution estimation. The conventional unimodal VAEs [3, 26] are optimized by maximizing the evidence lower bound (ELBO), which includes a reconstruction term and the Kullback-Leibler (KL) divergence term to measure the divergence from the variational posterior to the prior distribution of the latent variable. In a multimodal setting, the KL divergence is defined between the joint posterior and joint prior across the modalities, which is often estimated by the product of experts (PoE) [27, 28] or the mixture of experts (MoE) [29]. Based on such a prerequisite, many works extend the basic MVAE definition, such as missing modality handling [25], latent space modeling [24, 29], effective divergence modeling [5], *etc.* However, previous MVAE based frameworks essentially focus on the multimodal image generation task. In this work, we bring the MVAE to the AVS task and propose the conditional version [26, 30] of MVAE with practical multimodal information constraints for segmentation.

**Mutual Information Estimation.** Mutual Information (MI) captures the nonlinear statistical dependencies between variables, acting as a measure of actual dependence [31]. Specifically, for a pair of random variables  $X$  and  $Y$ , their MI  $I(X; Y)$  is defined as the KL divergence of the joint distribution  $p_{(X,Y)}$  from the product of the marginal distributions  $p(X)$  and  $p(Y)$ , which measures the shared information between  $X$  and  $Y$ . Although mutual information is simple in formation, as the log density ratio between the joint distribution  $p_{(X,Y)}$  and product of marginals  $p(X) \otimes p(Y)$  is intractable, it is usually estimated [32–34] instead of computed directly, leading to both MI maximization with a lower bound [35, 36] and MI minimization with an upper bound [37]. The MI maximization is usually applied for effective self-supervised representation learning [38–42] for the unimodal data to guarantee reliable feature representation. For the multimodal tasks, when each modality of data contains partial information of the target, both MI maximization and minimization can be applied [43–51], where the former aims to explore the task-driven feature across the modalities, and the latter is de-

signed to explore the complementary information of different modalities. Extensive researches show that effective MI optimization can not only lead to informative representation learning [52–59] but also is beneficial for achieving adversarial robustness [60].

### 3. Explicit Conditional Multimodal Learning

We denote the input data of our used AVSbench dataset [1] is  $X = \{\{x_t^v\}_{t=1}^T, x^a\}$ , *i.e.* the visual  $\{x_t^v\}_{t=1}^T$  for  $T$  non-overlapping yet continuous frames, audio  $x^a$  of the current clip.  $y = \{y_t\}_{t=1}^T$  is the output, *i.e.* the segmentation maps (we omit  $t$  for clear presentation). The goal of audio-visual segmentation is to segment the objects from the video  $\{x_t^v\}_{t=1}^T$  that produce the sound shown in the audio  $x^a$ . We introduce an Explicit Conditional Multimodal Variational Auto-Encoder (ECMVAE) using Jensen-Shannon divergence (Sec. 3.1) via latent space factorization (Sec. 3.2) to effectively model the shared representation between the two modalities (Sec. 3.3 and Sec. 3.4) for audio-visual segmentation. The overview of the proposed ECMVAE is shown in Fig. 1.

#### 3.1. Prerequisite

**Conditional Variational Auto-encoder.** We begin our prerequisite with the definition of the conditional variational auto-encoder (CVAE) [3, 26], which contains a generative process and an inference process. The generative process is to draw the latent variable  $z$  from the prior distribution  $p_\theta(z|X)$  with a given  $X$ , and generate the output via  $p_\theta(y|X, z)$ , where in our case  $X$  and  $y$  in the following derivations are  $X = \{\{x_t^v\}_{t=1}^T, x^a\}$  and  $y = \{y_t\}_{t=1}^T$ , respectively. The inference process of CVAE aims to infer the informative values of the latent variable  $z$  given the observed data  $X$  and  $y$  by computing the posterior  $p_\theta(z|X, y)$ , which is intractable and usually approximated with the variational posterior  $q_\phi(z|X, y)$ .  $\theta$  and  $\phi$  are the parameters of the true posterior and the approximated variational posterior, respectively. CVAE is trained to find the optimal generation parameters  $\theta^*$  and inference parameters  $\phi^*$  following the maximum log-likelihood learning pipeline:

$$\{\theta^*, \phi^*\} = \arg \max_{\theta, \phi} \log p_\theta(y|X), \quad (1)$$

where the log-likelihood term is achieved as:

$$\begin{aligned} & \log p_\theta(y|X) \\ &= \underbrace{\mathbb{E}_{q_\phi(z|X, y)} \log p_\theta(y|X, z) - D_{KL}(q_\phi(z|X, y) \| p_\theta(z|X))}_{\text{ELBO}(X, y, \theta, \phi)} \\ &+ D_{KL}(q_\phi(z|X, y) \| p_\theta(z|X, y)). \end{aligned} \quad (2)$$

Please see the supplementary material for the complete derivation.

By Jensen’s inequality, the Kullback–Leibler (KL) divergence term ( $D_{KL}$ ) in Eq. (2) is always greater or equal to zero, thus maximizing  $\log p_\theta(y|X)$  can be achieved by maximizing the evidence lower bound  $\text{ELBO}(X, y, \theta, \phi)$ :

$$\begin{aligned} \{\theta^*, \phi^*\} &= \arg \max_{\theta, \phi} \log p_\theta(y|X) \\ &= \arg \max_{\theta, \phi} \text{ELBO}(X, y, \theta, \phi). \end{aligned} \quad (3)$$

With the reparameterization trick [3], the KL term in  $\text{ELBO}(X, y, \theta, \phi)$  can be solved in closed-form if both the prior  $p_\theta(z|X)$  and posterior  $q_\phi(z|X, y)$  are Gaussian.

**Multimodal Conditional Variational Auto-encoder.** For the unimodal setting, both  $p_\theta(z|X)$  and  $q_\phi(z|X, y)$  can be obtained via the reparameterization trick [3], leading to closed-form solution of the KL divergence as both  $p_\theta(z|X)$  and  $q_\phi(z|X, y)$  are Gaussian. For the multimodal AVS data, the joint posterior and joint prior need to be estimated before we perform the joint generation process. The conventional solution to model the joint distribution is through the product of experts (PoE) [27, 28] or the mixture of experts (MoE) [29]. For the former, the joint distribution is defined as the product of each individual expert, which is Gaussian when each expert is Gaussian, leading to closed form KL computation. However, for PoE, less accurate modeling of one expert will completely destroy the joint distribution modeling. Further, PoE shows limitations in modeling the unimodal contribution due to its multiplicative nature. The additive nature of the MoE makes it effective for the optimization of each individual expert. However, as no closed form exists for the KL term, importance sampling (IS) is usually needed, which is computationally less efficient.

**JS Divergence Instead of KL Divergence.** Although MoE is computationally less efficient compared with PoE, its individual modal contribution modeling is attractive. Based on MoE, the  $D_{KL}$  term within  $\text{ELBO}(X, y, \theta, \phi)$  of Eq. (2) is the lower bound of the weighted sum of individual KLs:

$$\begin{aligned} & D_{KL}(q_\phi(z|X, y) \| p_\theta(z|X)) \\ & \leq \sum_{k=1}^K \phi_k D_{KL}(q_{\phi_k}(z|x^k, y) \| p_\theta(z|x^k)). \end{aligned} \quad (4)$$

where  $K$  is the number of modalities, and  $\sum_k \phi_k = 1$ . Although Eq. (4) is effective in providing lower bound with individual modal’s distribution for ELBO in Eq. (2), no joint distribution is involved. Following [5], a dynamic prior  $f_K$  is introduced, which is the mixture of the involved arguments (individual priors and posteriors), *i.e.*  $f_K$  can be defined as the arithmetic means as in MoE.

With the non-negative nature of KL divergence and the definition of JS divergence, we obtain a new lower bound of

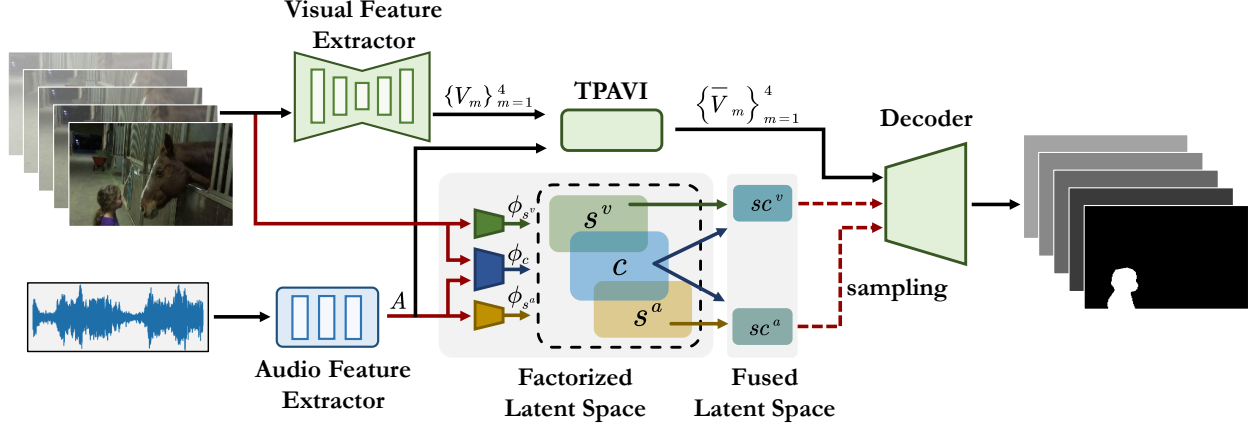


Figure 1. **Overview of the proposed ECMVAE for audio-visual segmentation.** The feature extractors are used to extract backbone features for the two modalities. We also design three latent encoders  $\phi_{s^v}$ ,  $\phi_{s^a}$ ,  $\phi_c$  to achieve latent space factorization and obtain both task-driven shared representation ( $c$ ) and modality-related specific representation ( $s^a$ ,  $s^v$ ), achieving explicit multimodal representation learning. The decoder is introduced to obtain the final segmentation maps, indicating the sound producers of the audio-visual data.

ELBO in Eq. (2) as:

$$\begin{aligned} \widehat{\text{ELBO}}(X, y, \theta, \phi) &\geq \mathbb{E}_{q_\phi(z|X, y)} \log p_\theta(y|X, z) \\ &- \sum_{k_1=1}^K \pi_{k_1} D_{KL}(q_\phi(z|X, y) \| f_K) - \sum_{k_2=1}^K \pi_{k_2} D_{KL}(p_\theta(z|X) \| f_K) \\ &= \underbrace{\mathbb{E}_{q_\phi(z|X, y)} \log p_\theta(y|X, z) - \text{JSD}(q_\phi(z|X, y), p_\theta(z|X))}_{\widehat{\text{ELBO}}(X, y, \theta, \phi)}, \end{aligned} \quad (5)$$

where  $\sum_{k=1}^{2K} \pi_k = 1$ , and JSD represents JS divergence. Eq. (5) provides lower bound of ELBO in Eq. (2), namely  $\widehat{\text{ELBO}}(X, y, \theta, \phi)$ , which is proven more stable for training [5], and robust to noise.

### 3.2. Latent Space Factorization

Besides stable training and noise robustness, we are also interested in modeling both shared and specific information of the audio-visual input (see Fig. 1) to fully explore their contribution. For each pair of example  $(x^k, y)$ , where  $x^k \in \{\{x_t^v\}_{t=1}^T, x^a\}$  indexes the modalities with  $K = 2$  in this paper, we factorize the latent space  $z$  into a modality-shared latent code  $c$  and modality-specific latent codes  $s^a$ ,  $s^v$ . Then  $\widehat{\text{ELBO}}(X, y, \theta, \phi)$  is re-defined as:

$$\begin{aligned} \widehat{\text{ELBO}}(X, y, \theta, \phi) &= \sum_{k=1}^K \mathbb{E}_{q_{\phi_c}(c|X, y)} \left[ \mathbb{E}_{q_{\phi_{s^k}}(s^k|x^k, y)} \left[ \log p_\theta(y|x^k, s^k, c) \right] \right] \\ &- \beta \sum_{k=1}^K D_{KL}(q_{\phi_{s^k}}(s^k|x^k, y) \| p_\theta(s^k|x^k)) \\ &- \beta \text{JSD}(q_{\phi_c}(c|X, y), p_\theta(c|X)), \end{aligned} \quad (6)$$

Please refer to the supplementary material for detailed derivation.

where  $q_{\phi_c}(c|X, y)$  and  $p_\theta(c|X)$  are the posterior and prior distributions of the shared representation.  $q_{\phi_{s^k}}(s^k|x^k, y)$  and  $p_{\theta_{s^k}}(s^k|x^k)$  are the posterior and prior distribution of the modality-specific latent codes.  $p_\theta(y|x^k, s^k, c)$  is the prediction generation model. All these models can be parameterized by deep neural networks and optimized via stochastic gradient descent. The hyper-parameter  $\beta = 0.1$  is introduced to achieve stable learning [61].

**Efficient Sampling.** The Eq. (6) shows that the generation process involves sampling from the joint shared posterior  $q_{\phi_c}(c|X, y)$  and posterior of modality-specific latent code  $q_{\phi_{s^k}}(s^k|x^k, y)$  of each modality, which is time-consuming. In practice, we first perform shared-specific representation fusion, and then we sample latent code from each fused space, achieving efficient sampling. Specifically, given the posterior of the latent codes  $c \sim q_{\phi_c}(c|X, y) \in \mathbb{R}^{T \times L}$ ,  $s^a \sim q_{\phi_{s^a}}(s^a|x^a, y) \in \mathbb{R}^{T \times L}$ ,  $s^v \sim q_{\phi_{s^v}}(s^v|x^v, y) \in \mathbb{R}^{T \times L}$  ( $L$  is the dimension of the latent space), we concatenate the shared representation with each specific representation to get the fused representation of each modality. Then we obtain  $sc^a, sc^v$ , representing the fused feature of audio and visual data, respectively. To achieve the reconstruction of  $p_\theta(y|x^k, s^k, c)$  in Eq. (6), instead of performing sampling from each specific latent code and shared latent code, we sample from  $sc^a, sc^v$ , and rewrite the reconstruction term, *i.e.* the first term in Eq. (6), as:

$$\mathcal{L}_{\text{rec}} = \sum_{sc \in \{sc^a, sc^v\}} [\mathbb{E} [\log p_\theta(y|x, sc)]] , \quad (7)$$

where  $x$  corresponds to the modality of data, *i.e.* audio or visual, of the current fused latent code  $sc$ .

**Hybrid Loss.** As the VAE samples from the posterior for training and prior to testing. To achieve consistent training



and testing, we define a Gaussian stochastic neural network (GSNN) [26] based objective by sampling from the prior distribution as well to avoid the posterior/prior distribution gap, leading to the hybrid objective as:

$$\widehat{\text{HELBO}}(X, y, \theta, \phi) = \alpha_1 \widehat{\text{ELBO}}(X, y, \theta, \phi) + (1 - \alpha_1) \mathcal{L}_{\text{GSNN}}, \quad (8)$$

where  $\alpha_1 = 0.5$  is used to balance the two objectives,  $\mathcal{L}_{\text{GSNN}}$  represents the reconstruction term of Eq. (6), which is achieved by sampling from the prior distribution.

The latent space factorization is effective in generating modality-shared and specific representations. However, no constraints are applied to the representations, making it hard to decide the reliability of the latent codes. We tackle this issue by proposing a representation orthogonality constraint in Sec. 3.3 and a shared information completeness regularization in Sec. 3.4.

### 3.3. Representation Orthogonality Constraint

We introduce a representation orthogonality constraint to ensure that the modality-shared and modality-specific representations capture different information within each uni-modal data. Specifically, given the latent codes  $c$ ,  $s^a$ ,  $s^v$ , we introduce the difference loss [62] as:

$$\mathcal{L}_{\text{diff}} = \|c^T s^a\|_F^2 + \|c^T s^v\|_F^2 + \|(s^a)^T s^v\|_F^2, \quad (9)$$

where  $\|\cdot\|_F^2$  is the squared Frobenius norm. With the difference loss, we aim to obtain the exclusive feature in each factorized feature representation.

### 3.4. Shared-Information Completeness

As discussed in Sec. 3.2, we fuse the shared representation with each modality-specific representation and obtain  $sc^a$ ,  $sc^v$ , representing the task-related information from audio and visual, respectively. To explicitly model the effectiveness of the fused latent space, we treat them ( $sc^a$ ,  $sc^v$ ) as two different views of the same target following representation learning [57, 58, 63], and introduce mutual information maximization as a regularizer to measure the shared information between  $sc^a$  and  $sc^v$ . Given two random variables  $SC^a$ ,  $SC^v$ , the mutual information is defined as:

$$I(SC^a; SC^v) = \mathbb{E}_{p(sc^a, sc^v)} \left[ \log \frac{p(sc^a, sc^v)}{p(sc^a) \cdot p(sc^v)} \right]. \quad (10)$$

According to Bayesian's law, we obtain the mutual information variational lower bound [36], namely  $I_{\text{ba}}$  via:

$$I(SC^a; SC^v) \geq \mathbb{E}_{p(sc^a, sc^v)} [\log q_\kappa(sc^a | sc^v)] + H(SC^a) \triangleq I_{\text{ba}}, \quad (11)$$

where  $q_\kappa(sc^a | sc^v)$  is the variational approximation of  $p(sc^a | sc^v)$ . Following [37],  $q_\kappa(sc^a | sc^v)$  is formulated as a multivariate Gaussian distribution  $q_\kappa(sc^a | sc^v) =$

$\mathcal{N}(sc^a | \mu(sc^v), \sigma^2(sc^v) \mathbf{I})$  to predict mean  $\mu(sc^v)$  and variance  $\sigma^2(sc^v)$ , respectively, where each statistic is modeled with two fully connected layers with Tanh activation function in the middle, and  $\kappa$  represent parameters of the four fully connected layers.  $H(SC^a)$  is the differential entropy of  $SC^a$ . As the audio encoder is fixed in this paper, we choose to simplify the entropy computation and treat  $H(SC^a)$  as a constant [64].

Based on the variational lower bound  $I_{\text{ba}}$ , we then define the shared-information completeness loss function as:  $\mathcal{L}_{\text{sic}} = -I_{\text{ba}}$ . Further, we introduce the hybrid loss function for the posterior and the prior distribution, leading to:

$$\mathcal{L}_{\text{sic}} = -\alpha_2 I_{\text{ba}}^{\text{po}} - (1 - \alpha_2) I_{\text{ba}}^{\text{pr}}, \quad (12)$$

where  $I_{\text{ba}}^{\text{po}}$  ( $I_{\text{ba}}^{\text{pr}}$ ) is the lower bound of the mutual information for the posterior (prior) distribution, and  $\alpha_2 = 0.5$  is introduced to balance the two objectives.

### 3.5. The Model

Four central modules or constraints are included in our framework (see Fig. 1), namely: **1**) “modal encoding” to extract the feature of each modality; **2**) “latent space encoding” for multimodal latent feature representation learning; **3**) “decoder” for the segmentation maps prediction; **4**) “objective function” for supervised learning and explicit multimodal representation constraints.

**Modal Encoding.** We perform two branches with two encoders to encode the visual and audio data. For the visual branch, we use the ImageNet pre-trained backbone followed by a one-layer convolution as neck to produce the multi-scale visual features  $\{V_m\}_{m=1}^4 \in \mathbb{R}^{T \times h_m \times w_m \times C_m}$ , where  $(h_m, w_m) = (H, W)/2^{m+1}$ ,  $C_m = 128$ .  $H, W$  is the spatial resolution of the input video. We use PVTv2 [65] or ResNet50 [66] as our visual backbone, which keeps the same as AVSBench [1]. For the audio branch, we follow [1], and use a frozen VGGish [67] model pre-trained on AudioSet [68] to process the spectrogram of input audio to extract audio features  $A \in \mathbb{R}^{T \times d}$ , where  $d = 128$ . And  $T = 5$  denotes the length of the video. We also keep the temporal pixel-wise audio-visual interaction (TPAVI) [1] module in our framework, which is a cross-modal attention based fusion module that takes visual features as query and value, audio features as key to achieve multi-scale feature fusion in the feature space and obtain  $\{\tilde{V}_m\}_{m=1}^4$ .

**Latent Space Encoding.** The main idea of our proposed method is achieving modality encoding on a reliable latent space (Sec. 3.2), as shown in Fig. 1. We use  $\phi_{s^v}$ ,  $\phi_{s^a}$ ,  $\phi_c$  parameterized by three simple neural networks to get latent feature embedding  $s^v, s^a, c \in \mathbb{R}^{T \times L}$  as prior distributions ( $L = 16$ ). For  $\phi_{s^v}$ , we use five convolutional layers followed by leakyReLU [69] with two fully connected layers to encode the input video sequence. While for  $\phi_{s^a}$ , we employ two fully connected layers to map the VGGish encoded

audio features into the latent space. Further, video sequence and audio features are fed into  $\phi_c$  jointly, thus we perform late fusion on the audio features and five convolutional layers followed by leakyReLU encoded visual features to obtain the joint distribution from the fused features. For the posterior network, we design three networks with the same structure and take the segmentation maps  $y$  as input by concatenating the video sequence and segmentation maps. We omit drawing the posterior space in Fig. 1 for easy viewing.

**Decoder.** We adopt the decoder of Panoptic-FPN [70] to decode the final segmentation maps for its flexibility and effectiveness, which is the same as AVSBench [1]. We expand  $sc^a, sc^v$  to feature map of the same spatial size as  $\bar{V}_4$  by adding two-dimensional gaussian noise with the tile operation, “sampling” is used to indicate this process in Fig. 1. The decoder takes both the deterministic features  $\{\bar{V}_m\}_{m=1}^4$  produced by the TPAVI [1] module and the expanded latent codes  $sc^a, sc^v$  from the fused latent space as input to produce the final segmentation maps.

**Objective Function.** As discussed above, our final objective function contains the optimization of the evidence lower bound and the practical constraints for latent space representation, and it can be defined as,

$$\mathcal{L} = -\widehat{\text{HELBO}}(X, y, \theta, \phi) + \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{sic}} + \lambda_3 \mathcal{L}_{\text{AVM}}, \quad (13)$$

where  $\widehat{\text{HELBO}}(X, y, \theta, \phi)$  indicates the lower bound for our proposed ECMVAE optimization, which is defined in detail in Eq. (8) and Eq. (6). We use a weighted structure-aware function [71] to compute the hybrid reconstruction part in  $\widehat{\text{HELBO}}(X, y, \theta, \phi)$ .  $\mathcal{L}_{\text{diff}}$  and  $\mathcal{L}_{\text{sic}}$  are the orthogonality constraint and the shared-information completeness loss defined in Eq. (9) and Eq. (12).  $\mathcal{L}_{\text{AVM}}$  indicates the audio-visual mapping loss proposed by [1] as a regularization term to promote the similarity between the audio-visual features. Empirically, we set the hyper-parameters  $\{\lambda_1, \lambda_2, \lambda_3\}$  as  $\{0.001, 0.01, 0.5\}$  for balanced training.

## 4. Experimental Results

### 4.1. Implementation Details

**Datasets.** We conduct experiments on the AVSBench [1] dataset, which contains 5,356 video sequences with corresponding audio data and binary per-pixel annotations. Each video in the dataset contains five frames, extracted separately from a five-second video, where the audio length is also five seconds. This dataset contains two settings, named S4 and MS3, for semi-supervised Single Sound Source Segmentation with only the first frame labeled, and fully supervised Multiple Sound Source Segmentation with all frames labeled. The evaluation is done for the entire five frames of the video under both S4 and MS3 settings on the test set.

Table 1. **Quantitative results on the AVSBench dataset [1]** in terms of mIoU and F-score under S4 and MS3 settings. We both report the performance with R50 and PVT as a backbone for the results of AVSBench [1] and Ours.

	Methods	S4		MS3	
		mIoU	F-score	mIoU	F-score
VOS	3DC [72]	57.10	0.759	36.92	0.503
	SST [73]	66.29	0.801	42.57	0.572
SOD	iGAN [74]	61.59	0.778	42.89	0.544
	LGVT [75]	74.94	0.873	40.71	0.593
AVS	AVSBench (R50) [1]	72.79	0.848	47.88	0.578
	AVSBench (PVT) [1]	78.74	0.879	54.00	0.645
	Ours (R50)	76.33	0.865	48.69	0.607
	Ours (PVT)	<b>81.74</b>	<b>0.901</b>	<b>57.84</b>	<b>0.708</b>

**Training Details.** We conduct experiments on Pytorch [76] with a single NVIDIA A100 GPU. The Adam [77] solver is used to optimize our network with a learning rate of  $1 \times 10^{-4}$ . The training batch size is set to 4. We train the network on the S4 subset for 15 epochs, and on the MS3 subset for 30 epochs. For the MS3 setting, we use all ground-truth of the five frames to build the posterior latent space of our model. While for S4, we repeat the ground-truth of the first frame five times to build the posterior latent space.

### 4.2. Comparison with Baseline Methods

**Quantitative Comparison.** Follow the comparison settings with AVSBench [1], we compare the performance of our ECMVAE with baseline AVS models and other related tasks, including video object segmentation (VOS) and salient object detection (SOD). The performance on Mean Intersection over Union (mIoU) and F-score is reported in Table 1. It can be observed that our method consistently achieves significantly superior segmentation performance than the state-of-the-art methods, especially with 3.00 and 3.84 higher mIoU than the previous AVS method [1], at the S4 and MS3 settings with PVTv2 (“PVT”) backbone. There is also a consistent performance improvement with ResNet (“R50”) backbone. The performance gain comes from our designed multimodal VAE with explicit constraints for representation learning. Our method also significantly outperforms the VOS and SOD methods, demonstrating the addition of audio information to the segmentation performance.

**Qualitative Comparison.** We provide a qualitative comparison between our proposed method and [1] in Fig. 2. Our proposed ECMVAE provides a better audio temporal and spatial localization quality, leading to better segmentation performance, especially for the left samples, in localization of the *piano keys*, which is not salient but producing a sound in this scene. Our method also achieves better segmentation performance for background noise handling and richer foreground details in the right samples in Fig. 2.

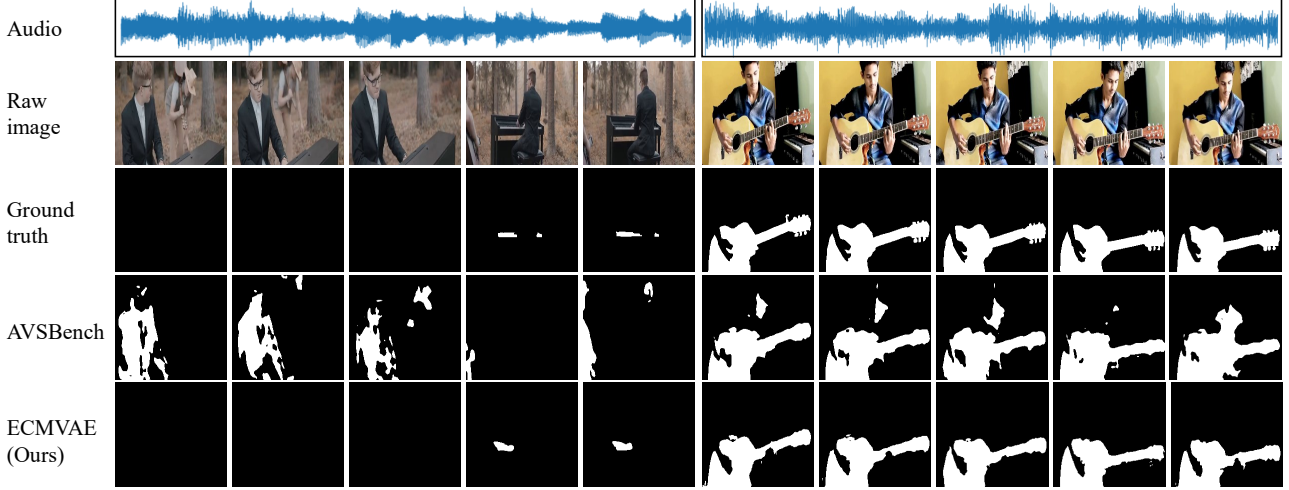


Figure 2. **Qualitative comparison** between our proposed ECMVAE and AVSBench [1]. Our method competently achieves high segmentation performance with better audio temporal and spatial localization quality and detail handling.

### 4.3. Ablation Studies

We conduct ablation studies of our proposed method. All variations are trained with the PVT backbone.

Table 2. **Ablation of the VAE based multimodal learning.** We implement a “CVAE” without audio and “CMVAE” with audio-visual joint distribution estimation.

Methods	S4		MS3	
	mIoU	F-score	mIoU	F-score
[1] w/o audio	77.80	-	48.20	-
CVAE	78.12	0.878	49.26	0.643
CMVAE	<b>80.05</b>	<b>0.889</b>	<b>54.99</b>	<b>0.653</b>

**Multimodal VAE.** We explore the effectiveness of the Multimodal VAE in Table 2. Firstly, we remove the audio part of the model and disable the TPAVI module to explore the importance of the audio information, leading to a simple unimodal CVAE [30] framework with only video input, which is denoted as “CVAE”. For comparison, we implement a “CMVAE” using the audio signal but without the latent space factorization and our proposed constraints. The better performance of “CMVAE” compared with “CVAE” indicates the importance of audio for AVS, especially when multiple sound sources exist. We also brought the ablation result of the model without the audio input from [1] and compare it with “CVAE”. The results show that in the absence of audio signals, the VAE structure can still improve segmentation performance, due to the ability of VAE to model the latent space of visual features.

**Latent Space Factorization.** As described in Sec. 3.2, we factorize the latent space of multimodal VAE into modality-shared ( $c$ ) and modality-specific ( $s^v$  and  $s^a$ ) representations. As reported in Table 3, removing the latent space

Table 3. **Ablation of the latent space factorization.** “Model” indicates using VAE for latent factorization or using AE for feature factorization. “Factor.” denotes whether using factorization. “dim.” represents the size of the latent dimensions.

Model	Factor.	dim.	S4		MS3	
			mIoU	F-score	mIoU	F-score
VAE	-	16	80.05	0.889	54.99	0.653
	-	48	80.13	0.890	55.09	0.657
	✓	16	<b>80.78</b>	<b>0.893</b>	<b>56.38</b>	<b>0.676</b>
AE	-	-	78.74	0.879	54.00	0.645
	✓	-	78.92	0.881	54.82	0.651

factorization leads to obvious performance degradation. We also train a non-factorized model with  $2\times$  larger latent dimensions, which holds comparable latent space capacities with the factorized model. It can be seen that the performance gain from the larger latent space dimensions is not as obvious as the factorization strategy. Further, we compare feature factorization (“AE”) on the feature space of [1] with our proposed latent space factorization (“VAE”). Table 3 shows that only performing factorization on a semantic meaningful and continuous latent space, *i.e.* via using VAE [3], can achieve larger performance improvements.

**JS Divergence.** We conduct experiments of PoE and MoE with KL divergence to show the trade-off of JS divergence between the computational efficiency of the inference process and the predictive quality of the generation process. The performance in Table 4 shows the effectiveness of JS divergence on both S4 and MS3 settings. Note that, “PoE”, “MoE” and “JS” are based on our formulation of the task, which has not been explored in the AVS area.

**Orthogonality Constraint.** As reported in Table 5, the or-

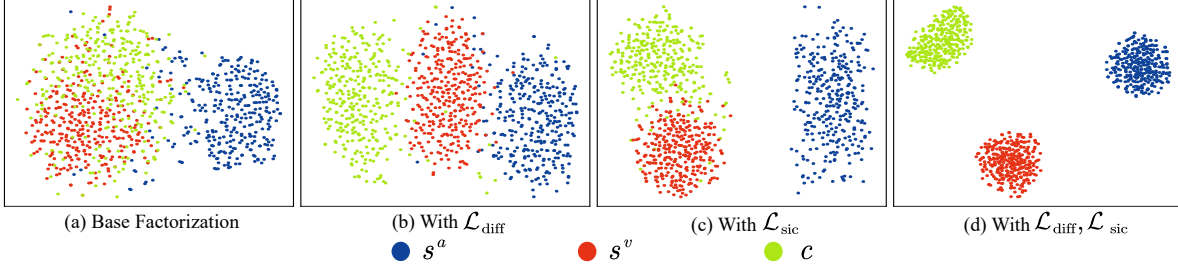


Figure 3. **Visualization of the modality-shared and modality-specific latent codes** ( $s^v$ ,  $s^a$ ,  $c$ ) in the MS3 testing set using t-SNE [78] projections. Best viewed on screen.

Table 4. **Ablation of the JS Divergence.** We implement “PoE” and “MoE” with KL divergence, and a simple “KL” model.

Methods	S4		MS3	
	mIoU	F-score	mIoU	F-score
KL	80.78	0.893	56.38	0.676
PoE	81.38	0.894	57.35	0.688
MoE	81.49	0.897	57.53	0.694
JS	<b>81.74</b>	<b>0.901</b>	<b>57.84</b>	<b>0.708</b>

Table 5. **Ablation of the latent space constraints.**  $\mathcal{L}_{\text{diff}}$ ,  $\mathcal{L}_{\text{sic}}$  indicate our proposed loss functions for multimodal learning.

$\mathcal{L}_{\text{diff}}$	$\mathcal{L}_{\text{sic}}$	S4		MS3	
		mIoU	F-score	mIoU	F-score
-	-	81.09	0.895	57.01	0.684
✓	-	81.51	0.899	57.65	0.692
-	✓	81.47	0.898	57.51	0.694
✓	✓	<b>81.74</b>	<b>0.901</b>	<b>57.84</b>	<b>0.708</b>

thogonality constraint provides 0.64 mIoU gain under the MS3 setting, which facilitates the latent space factorization. We also perform t-SNE [78] projection to visualize the factorized latent code with and without such constraint. As compared between Fig. 3 (a) and (b), the three latent codes in the latent space are divided into different subspaces to ensure that each latent code encodes different information.

**Mutual Information Maximization.** As compared in Table 5, the mutual information maximization by  $\mathcal{L}_{\text{sic}}$  also serves a crucial impact for the explicit constraint of the shared latent space and improves the mIoU from 57.51 to 57.84. Since the  $\mathcal{L}_{\text{sic}}$  maximizes the mutual information between  $s^v$  and  $s^a$ , which are fused from the factorization latent codes  $s^v$ ,  $s^a$ ,  $c$ . This increases the amount of information contained in latent space and makes it more effective for facilitating factorization. Fig. 3 also confirms this view and demonstrates the effectiveness of  $\mathcal{L}_{\text{sic}}$  in achieving effective “multi-view” representation learning [57, 58, 63].

#### 4.4. Analysis

**Pre-training on the Single-source Subset.** In AVS-Bench [1], they conducted experiments by initializing

model parameters by pre-training on S4 dataset. We also perform experiments with such setting (see Table 6). We can observe that both our method and AVSBench [1] can benefit from the model pre-trained on S4. The PVT-based model can gain 2.97% mIoU performance by such a strategy and reach 60.81% mIoU. The pre-training strategy can bring more significant improvements (8.87% mIoU) to ResNet50-based models and reach 57.56% mIoU, which is even higher than PVT-based models (57.34%).

Table 6. **Performance comparison with different initialization strategies for MS3 dataset.** As AVSBench [1] does not report its F-score in the paper, we only report its mIoU. The values in parentheses indicate the performance improvement based on S4 pre-training compared with training from scratch.

Methods	From scratch		Pre-trained on S4	
	mIoU	F-score	mIoU	F-score
AVSBench (R50) [1]	47.88	-	54.33 (↑ 6.45)	-
AVSBench (PVT) [1]	54.00	-	57.34 (↑ 3.34)	-
Ours (R50)	48.69	0.607	57.56 (↑ 8.87)	0.674
Ours (PVT)	<b>57.84</b>	<b>0.708</b>	<b>60.81</b> (↑ 2.97)	<b>0.729</b>

Table 7. **Parameters and inference time.**

Methods	R50		PVT	
	Param. (M)	Time (ms)	Param. (M)	Time (ms)
AVSBench [1]	70.50	28	101.32	53
Ours	<b>33.97</b>	<b>23</b>	<b>91.18</b>	<b>46</b>

**Parameters and Efficiency.** In Table 7, we compare the parameters and inference time between ours and AVSBench. Note that although posterior nets and prior nets are used in our framework, as all the latent space encoders are quite lightweight (1M), thus our model capacity will not change significantly. Moreover, we replace the neck from ASPP [79] to one-layer convolution and reduce the number of neck channels from 256 to 128, which leads to smaller parameter numbers and faster inference speed.

**Limitations.** Similar to the other VAE [3] based solutions, our model also suffers from the risk of posterior collapse, where the posterior of the latent variable is equal to



prior [80,81], making  $y$  in our case not encoded in the latent variables. To avoid such phenomenon, contrastive learning [82, 83] can be studied to learn more compact features of each modality or score based diffusion models [84–87] can be investigated for more informative latent space.

## 5. Conclusion

We have worked on audio-visual segmentation (AVS), aiming to segment the sound producers of the scene. As audio data can be treated as “command”, we argue extensive exploration of audio is critical for effective AVS. Inspired by this observation, we have introduced an Explicit Conditional Multimodal Variational Auto-Encoder (ECM-VAE) for an audio-visual segmentation model with latent space factorization with explicit constraints to extensively explore the shared and specific representations of audio and visual data. Extensive experimental results verify the effectiveness of our proposed framework. Although we focused on AVS with two modalities, the proposed framework can be extended to more modalities and other audio-visual collaboration tasks [7, 14, 15].

## 6. Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (62271410), and the Fundamental Research Funds for the Central Universities.

## References

- [1] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 5, 6, 7, 8
- [2] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 1
- [3] Diederik Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 1, 2, 3, 7, 8
- [4] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2
- [5] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3, 4
- [6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [7] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 9
- [8] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision (ECCV)*, 2018. 2
- [9] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019. 2
- [10] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Asian Conference on Computer Vision (ACCV)*, 2020. 2
- [11] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, 2018. 2
- [12] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [13] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2
- [14] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: multimodal pyramid attentional network for audio-visual event localization and video parsing. In *ACM International Conference on Multimedia (MM)*, 2022. 2, 9
- [15] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 9
- [16] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [17] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [18] Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2
- [19] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision (ECCV)*, 2020. 2

- [20] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 2
- [21] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [22] Tom Joy, Yuge Shi, Philip Torr, Tom Rainforth, Sebastian M Schmon, and Siddharth N. Learning multimodal VAEs through mutual supervision. In *International Conference on Learning Representations (ICLR)*, 2022. 2
- [23] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016. 2
- [24] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning (ICML)*, 2019. 2
- [25] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [26] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 2, 3, 5
- [27] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002. 2, 3
- [28] Yanshuai Cao and David J Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014. 2, 3
- [29] Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3
- [30] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 7
- [31] Justin B. Kinney and Gurinder S. Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, pages 3354–3359, 2014. 2
- [32] Arnab Mondal, Arnab Bhattacharjee, Sudipto Mukherjee, Himanshu Asnani, Sreeram Kannan, and Prathosh A P. C-mi-gan : Estimation of conditional mutual information using minmax formulation. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020. 2
- [33] Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi : Classifier based conditional mutual information estimation. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020. 2
- [34] Alan Yang, AmirEmad Ghassami, Maxim Raginsky, Negar Kiyavash, and Elyse Rosenbaum. Model-augmented conditional mutual information estimation for feature selection. In *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020. 2
- [35] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning (ICML)*, pages 5171–5180, 2019. 2
- [36] David Barber and Felix V. Agakov. The im algorithm: A variational approach to information maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 201–208, 2003. 2, 5
- [37] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning (ICML)*, 2020. 2, 5
- [38] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Orntner. Learning disentangled representations via mutual information estimation. In *European Conference on Computer Vision (ECCV)*, 2020. 2
- [39] Detai Xin, Tatsuya Komatsu, Shinnosuke Takamichi, and Hiroshi Saruwatari. Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual tts. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021. 2
- [40] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [41] Sobhan Soleymani, Ali Dabouei, Fariborz Taherkhani, Jeremy Dawson, and Nasser M Nasrabadi. Mutual information maximization on disentangled representations for differential morph detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2
- [42] Yijun Xiao and William Yang Wang. Disentangled representation learning with wasserstein total correlation. *arXiv preprint arXiv:1912.12818*, 2019. 2
- [43] Yiqiao Mao, Xiaoqiang Yan, Qiang Guo, and Yangdong Ye. Deep mutual information maximin for cross-modal clustering. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [44] Ruizhi Liao, Daniel Moyer, Miriam Cha, Keegan Quigley, Seth Berkowitz, Steven Horng, Polina Golland, and William M Wells. Multimodal representation learning via maximization of local mutual information. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 273–283, 2021. 2

- [45] Zhenhong Zou, Linhao Zhao, Xinyu Zhang, Zhiwei Li, Dafeng Jin, and Tao Luo. Mimi: Mutual information-driven multimodal fusion. In *Cognitive Systems and Signal Processing*, pages 142–150, 2021. 2
- [46] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloé Clavel. Improving multimodal fusion via mutual dependency maximisation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 231–245, 2021. 2
- [47] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [48] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations (ICLR)*, 2019. 2
- [49] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9180–9192, 2021. 2
- [50] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, pages 1692–1700, 2021. 2
- [51] Imant Daunhawer, Thomas M. Sutter, Ričards Marcinkevičs, and Julia E. Vogt. Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR*, 2020. 2
- [52] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 3
- [53] Youssef Mroueh, Igor Melnyk, Pierre Dognin, Jarret Ross, and Tom Sercu. Improved mutual information estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 3
- [54] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, pages 531–540, 2018. 3
- [55] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3
- [56] Mete Kemertas, Leila Pishdad, Konstantinos G. Derpanis, and Afsaneh Fazly. Rankmi: A mutual information maximizing ranking loss. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [57] Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3, 5, 8
- [58] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3, 5, 8
- [59] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations (ICLR)*, 2019. 3
- [60] Dawei Zhou, Nannan Wang, Xinbo Gao, Bo Han, Xiaoyu Wang, Yibing Zhan, and Tongliang Liu. Improving adversarial robustness via mutual information estimation. In *International Conference on Machine Learning (ICML)*, 2022. 3
- [61] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. 4
- [62] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 5
- [63] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J. Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12793–12802, June 2021. 5, 8
- [64] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 5
- [65] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 2022. 5
- [66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5
- [67] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017. 5
- [68] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal,

- and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017. 5
- [69] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, 2013. 5
- [70] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6
- [71] Jun Wei, Shuhui Wang, and Qingming Huang. F<sup>3</sup>net: fusion, feedback and focus for salient object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 6
- [72] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. *arXiv preprint arXiv:2008.11516*, 2020. 6
- [73] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6
- [74] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. Generative transformer for accurate and reliable salient object detection. *arXiv preprint arXiv:2104.10127*, 2021. 6
- [75] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6
- [76] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6
- [77] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6
- [78] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8
- [79] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 8
- [80] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015. 9
- [81] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 9
- [82] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 9
- [83] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 9
- [84] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015. 9
- [85] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 9
- [86] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11895–11907, 2019. 9
- [87] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 9