LEARNING AUDIO-VISUAL CORRELATIONS FROM VARIATIONAL CROSS-MODAL GENERATION

Ye Zhu[1] *Yu Wu*[2] *Hugo Latapie*[3] *Yi Yang*[2] *Yan Yan*[1]

[1] Illinois Institute of Technology, USA
[2] ReLER, University of Technology Sydney, Australia
[3] Cisco, USA

ABSTRACT

People can easily imagine the potential sound while seeing an event. This natural synchronization between audio and visual signals reveals their intrinsic correlations. To this end, we propose to learn the audio-visual correlations from the perspective of cross-modal generation in a self-supervised manner, the learned correlations can be then readily applied in multiple downstream tasks such as the audio-visual cross-modal localization and retrieval. We introduce a novel Variational AutoEncoder (*VAE*) framework that consists of Multiple encoders and a Shared decoder (*MS-VAE*) with an additional Wasserstein distance constraint to tackle the problem. Extensive experiments demonstrate that the optimized latent representation of the proposed *MS-VAE* can effectively learn the audio-visual correlations and can be readily applied in multiple audio-visual downstream tasks to achieve competitive performance even without any given label information during training.

Index Terms— Audio-visual correlations, Variational autoencoder, Cross-modal generation.

1. INTRODUCTION

As humans, we can naturally imagine the possible visual frames while hearing the corresponding sound or imagine the potential sound while seeing an event happening. Therefore, the correlations between audio and visual information accompanying an event can be modeled in the perspective of generation, *i.e.*, the corresponding audio signals and visual frames can generate each other. Moreover, the natural correspondence between audio and visual information from the videos makes it possible to accomplish this objective in a self-supervised manner without additional annotations.

Audio and visual perceptions are both essential sources of information for humans to explore the world. Audio-visual cross-modal learning has thus become a research focus in recent years [1, 2, 3, 4, 5, 6, 7, 8, 9]. The correlations between audio and visual signals are the key in this field. Recent studies in the audio-visual cross-modal field largely focus on representation learning that incorporates the information from both modalities in a discriminative way, and then applying the learned feature embedding in relevant audio-visual tasks such as sound source localization [5, 10, 11], sound source separation [2, 3, 11], cross-modal retrievals [12, 3, 5, 13] and cross-modal localization [13, 14]. In contrast, another branch of research work exploits to learn the correlations in a generative manner [15, 6, 16]. Hu *et al.* [7] introduce Deep Multimodal Clustering for capturing the audio-visual correspondence. Korbar *et al.* [1] propose a cooperative learning schema to obtain multi-sensory representation from self-supervised synchronization. Arandjelovic and Zisserman [3] propose to learn a mutual feature embedding through the audio-visual correspondence task. As for more concrete audio-visual downstream tasks, Gao *et al.* [10] look into the problem of separating different sound sources based on a deep multi-instance multi-label network. Among those above studies with concrete downstream tasks, most of them learn the audio-visual correlations in a discriminative way to obtain better performance, which usually requires label information. Our work tackles the problem from a different perspective based on generative models via self-supervised learning.

Overall, we have two motivations to fulfill in this work: leveraging the label-free advantage to learn the intrinsic audio-visual correlations from cross-modal generations via the proposed *MS-VAE* framework, and achieving competitive performance for multiple audio-visual downstream tasks at the same time. Specifically, the proposed *MS-VAE* is a Variational AutoEncoder framework with Multiple encoders and a Shared decoder. VAE [17] is a popular class of generative models that synthesizes data with latent variables sampled from a variational distribution, generally modeled by Gaussian distributions. Based on the properties of VAE, we propose to use the latent variables to represent modality-specific data. Then these latent variables need to be aligned in order to complementarily present the happening event in two perspectives. Finally, the aligned latent variable should be able to generate both audio and visual information to construct the corresponding data pair. The optimized latent variable hence automatically learns the intrinsic audio-visual correlations during the process of cross-modal generation,

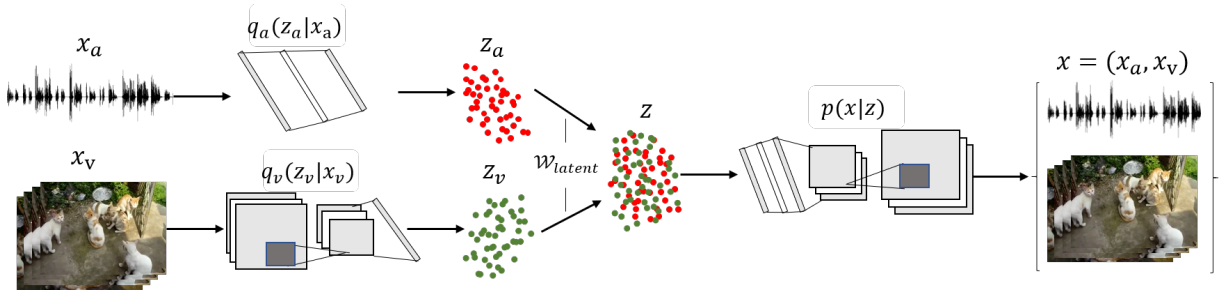


Fig. 1. Schematic overview of our proposed *MS-VAE* model.

and is ready to be directly applied in audio-visual tasks.

One practical challenge to apply the VAE framework in multi-modal learning is that VAE often suffers from the degeneration in balancing simple and complex distributions [18] due to the large dimensionality difference between audio and video data. We adopt a shared-decoder architecture to help avoiding the degeneration and to enforce the mutuality between audio and visual information. To further obtain a better alignment of the latent space, we derive a new evidence lower bound (ELBO) with a Wasserstein distance [19, 20, 21] constraint, which formulates our objective function.

The main contributions of our work can be summarized as follows: 1) We model the audio-visual correlations from the perspective of cross-modal generation. We make a shared-latent space assumption to apply a unified representation for two modalities, by deriving the objective function from a new lower bound with a Wasserstein distance constraint. 2) We propose the *MS-VAE* network, which is a novel self-supervised learning framework for the audio-visual cross-modal generation. *MS-VAE* generates a corresponding audio-video pair from either single modality input and alleviates the degeneration problem in VAE training. 3) The learned latent representations from *MS-VAE* can be readily applied in multiple audio-visual tasks. Experiments on AVE dataset [13] show that our unsupervised method is able to achieve performance comparable or superior to the supervised methods on the challenging localization and retrieval tasks, even trained without any labels.

2. METHODOLOGY

2.1. MS-VAE for Cross-Modal Learning

Our *MS-VAE* network is composed of separate encoders and one shared decoder. $q_{a,\phi_a}(z_a|x_a)$ is the encoder that encodes audio input data x_a into a latent space z_a, and $q_{v,\phi_v}(z_v|x_v)$ is the encoder for visual input data x_v to construct another latent space z_v. Ideally, we wish to obtain an aligned mutual latent space z where $z = z_a = z_v$. The shared decoder $p_\theta(x|z)$ aims to reconstruct the original data pair x from this mutual latent space z in training, in which case the expected reconstructed data should consist of x_a and x_v, we denote the pair of audio and video data as $x = (x_a, x_v)$. ϕ_a, ϕ_v and θ are the model parameters, which we omit in the following formulations to reduce redundancy.

The goal of our model resembles to the original VAE [17], where we target to maximize the log-probability $\log p(x)$ of the reconstruction data pair x from the desired mutual latent space z, i represents either the modality a (audio) or v (visual). This model design leads to a similar variational lower bound as the original VAE [17] as follows:

$$\log p(x) \geq \mathbb{E}_{z_i \sim q_i(z_i|x_i)}[\log p(x|z_i)] - \mathrm{KL}(q_i(z_i|x_i)||p(z_i)), \tag{1}$$

where KL denotes the Kullback-Leibler divergence, defined as $\mathrm{KL}(p(x)||q(x)) = \int_x p(x) \log \frac{p(x)}{q(x)}$, which measures the similarity between two distributions and is always positive. To build the relation between audio and video components, we rewrite Equation 1 as a mixture of log-likelihood conditioned on the latent variable from different modality. A Wasserstein distance loss [22], which we refer as $\mathcal{W}_{\text{latent}}$, is further added to better encourage the alignment between two latent space. Since the Wasserstein distance is always positive, the inequality remains valid. In this case, we obtain a new lower bound, whose equality is obtained only when the modeled distribution is the same as data distribution, as well as z_a and z_v are perfectly aligned:

$$\begin{aligned}\log\, p(x) \geq \mathbb{E}_z[\log(p(x|z)] - \frac{1}{2}[\mathrm{KL}(q_a(z|x_a)||p(z)) \\ + \mathrm{KL}(q_v(z|x_v)||p(z))] - \mathcal{W}_{\text{latent}}(q_a(z_a|x_a)||q_v(z_v|x_v)).\end{aligned} \tag{2}$$

2.2. Network and Training

The schematic overview of the proposed *MS-VAE* architecture is illustrated in Figure 1. We have separate encoders q_a and q_v for audio and visual inputs, a shared decoder p is used to generate the corresponding audio and visual data pair. Wasserstein distance is computed between two latent variables z_a and z_v to encourage the alignment between the latent space

using the similar approach as in WAE [23], where we sample from latent variables z_a and z_v to compute $\mathbb{E}_{q_a,q_v}\,[||z_a, z_v||_2]$.

The ultimate goal of the proposed *MS-VAE* is to obtain an aligned latent representation that automatically learns the audio-visual correlations. During training, for each epoch, we reconstruct the audio-visual pair from either audio or visual input. The encoder returns the μ_i and σ_i for the Gaussian distribution, z_i is sampled from $\mathcal{N}(\mu_i, \sigma_i)$. The reconstruction loss Mean Square Error (MSE) is computed between the reconstructed pair $\hat{x_i}$ from modality i and the input ground truth pair x. The total loss contains the reconstruction loss, KL divergence and the Wasserstein latent loss. No label information is given in the entire training process. Overall speaking, we have three loss terms to optimize:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{KL}} + \lambda_3 \mathcal{W}_{\text{latent}}. \tag{3}$$

Empirically, we choose λ_1 and λ_3 to be 1, λ_2 is set to be 0.1 in the first 10 training epochs and then reduced to 0.01 to encourage better reconstruction.

3. EXPERIMENTS

3.1. Dataset and Evaluation Metrics

We expect the data for our experiments have intrinsic correlations in describing an audio-visual event, we therefore adopt the AVE dataset. The AVE dataset [13] is a subset of AudioSet [24] that contains 4,143 videos labeled with 28 event categories. Each video has a duration of 10s, and at least 2s are labeled with audio-visual events, such as baby crying and dog barking *etc.*. Apart from the event categories labels for the entire video, each video is also temporally divided into 10 segments with audio-visual event boundaries, indicating whether the one-second segment is event-relevant or the background. The labels in *MS-VAE* are purely used for evaluation purposes in our experiments, no labels are used in training.

We mainly apply our proposed model in two downstream tasks: the cross-modal localization (CML) [13] and the audio-visual retrieval [3]. The CML task contains two subtasks, including localizing the visual event boundary from audio signals (A2V) and vice versa (V2A). The evaluation accuracy is computed based on the strict exact match, which means that we only count correct when the match location is exactly the same as its ground truth. For the retrieval task, the mean reciprocal rank (MRR) is used to evaluate the performance. MRR calculates the average value of the reciprocal of the rank at which the first relevant information is retrieved across queries. In addition to the concrete audio-visual tasks, we also include further qualitative ablation studies on the learned latent space to provide a more comprehensive analysis for the proposed *MS-VAE* model.

Setting	Method	A2V ↑	V2A ↑	Average ↑
Spv.	DCCA [27]	34.1	34.8	34.5
	AVDLN [13]	35.6	44.8	40.2
	AVSDN [25]	37.1	45.1	40.9
	AVFB [26]	37.3	46.0	41.6
Unspv.	Ours	25.0 ± 0.9	38.8 ± 0.4	31.9 ± 0.7
	Ours+$\mathcal{W}$	37.4 ± 1.2	40.0 ± 1.4	38.7 ± 1.3

Table 1. Quantitative evaluations on the CML task in terms of the exact matching accuracy. $Spv.$ means supervised and $Unspv.$ means unsupervised.

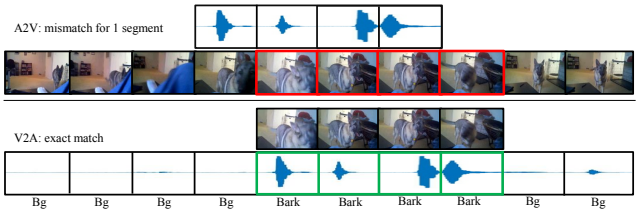


Fig. 2. Example of qualitative results for the CML task.

3.2. Cross-Modal Localization Task

Cross-modal localization is proposed in [13], in which we wish to temporally locate a given segment of one modality (audio/visual) data in the entire sequence of the other modality. The localization is realized in two directions, *i.e.*, visual localization from given audio segments (A2V) and audio localization from given visual segments (V2A). This task especially emphasizes the correlations between visual and audio signals since not only the correlations between different modalities are required, the temporal correlations are also needed to successfully fulfill the task [13, 14, 25, 26].

In inference, for the sub-task A2V, we adopt the sliding window strategy as in [13] to optimize the following objective: $t^* = \text{argmin}_t \sum_{s=1}^{l} \mathcal{D}_{cml}(V_{s+t-1}, \hat{A}_s)$, where $t^* \in \{1, ..., T - l + 1\}$ is the start time when audio and visual content synchronize, T is the total length of a testing video sequence, and l is the length of the audio query $\hat{A}_s$. The time position that minimize the cumulative distance between audio segments and visual segments is chosen to be the matching start position. For $\mathcal{D}_{cml}$ in our experiments, we compute two terms, *i.e.*, the Wasserstein latent distance $\mathcal{W}_{latent}$ between two latent variables encoded from audio and visual segments, and the Euclidean distance $\mathcal{D}_{gen}$ between the generated pair by audio and visual segment: $\mathcal{D}_{cml} = \mathcal{W}_{latent} + \mathcal{D}_{gen}$. Similarly for another sub-task V2A.

The quantitative results are shown in Table 1. We compare our method with other network models that learn the audio-visual correlation in a supervised manner. It is worth noting that our proposed self-supervised model achieves comparable performance with the supervised state-of-the-art methods. Figure 2 shows an example of a mismatch and exact match

Setting	Method	A-A ↑	A-V↑	V-A↑	V-V↑
Spv.	AVDLN [13]	0.34	0.17	0.20	0.26
	AVSDN [25]	0.38	0.19	0.21	0.25
	AVFB [26]	0.37	0.20	0.20	0.27
Unspv.	L^3 [28]	0.13	0.14	0.13	0.14
	AVE [3]	0.16	0.14	0.14	0.16
	Ours	0.24	0.14	0.15	0.27
	Ours+$\mathcal{W}$	0.37	0.17	0.18	0.24

Table 2. Cross-modal and intra-modal retrieval results in terms of MRR. The columns headers denote the modalities of the query and the database. For example, A-V means retrieve visual frames from audio query. *Spv.* and *Unspv.* denote supervised and unsupervised.

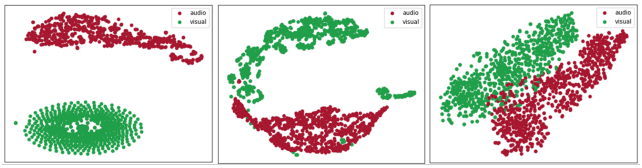
for the cross-modal localization task. In A2V sub-task, we are given an audio query accompanying an event, *e.g.,* dog bark, we want to localize its corresponding positions in the entire 10s visual frame sequence. Similarly in the V2A sub-task. Only the exact match is considered to be correct, thus making the task very challenging.

3.3. Cross-Modal Retrieval Task

Cross-modal retrieval task proposed in [3] seeks to find the relevant audio/visual segment given a single audio or visual query. As in [3], we randomly sample a single event-relevant pair of visual frame and audio signal from each video in the testing set of AVE dataset to form the retrieval database. The retrieval task contains intra-modal (*e.g.*, audio-to-audio) and cross-modal (*e.g.*, audio-to-visual) categories. Each item in the test set is used as a query. The identical item from the database for each given query in intra-modal retrieval is removed to avoid bias. All the labels are only used for evaluations to calculate the MRR metric, which computes the average of the reciprocal of the rank for the first correctly retrieved items (*i.e.*, items with the same category label as the query).

We fine-tune the L^3-Net [28] and AVE-Net [3] in the training set of AVE dataset. AVE-Net incorporates a Euclidean distance between two separate feature embeddings to enforce the alignment. Both models learn audio-visual embedding in a self-supervised manner. We use the same inference as presented in [28] and [3] for L^3-Net and AVE-Net, respectively. In addition to these two unsupervised baselines, we also adopt the supervised models to perform the retrieval tasks. Similar to the previous CML task, these models are trained to minimize the similarity distance between two embeddings of a corresponding audio-visual pair. In inference, we use a similar technique as in cross-modal localization task, which is to retrieve the sample that has the minimum distance score. For *MS-VAE*, we compare the sum of Wasserstein latent distance and the reconstruction error between the given query and all the test samples from the retrieval database.

Table 2 presents the quantitative comparison between the proposed *MS-VAE* and other methods. We achieve better performance in all four sub-tasks compared to the unsupervised baselines, and achieve competitive performance close to the supervised models. This experiment further proves the effectiveness of leveraging audio-visual correlations for *MS-VAE*.



(a) Latent space with CM-VAE (b) Latent space with MS-VAE (c) Latent space with MS-VAE + $\mathcal{W}$

Fig. 3. Latent space comparison. t-SNE visualization for testing segments of corresponding audio-visual pair.

3.4. Ablation Analysis

We perform additional ablation studies on the shared decoder and the Wasserstein restriction components to show that they are both significant factors that contribute to the learned audio-visual correlations. Figure 3 shows the qualitative results for the latent space comparison among *CM-VAE* [29], our *MS-VAE* and *MS-VAE + $\mathcal{W}$* in the form of latent space visualization. Specifically, *CM-VAE* is proposed in [29] for cross-modal generation for hand poses (*e.g.*, generating 3D hand poses from RGB images). *CM-VAE* adopts separate encoders and decoders for each modality, and can be considered as *MS-VAE* without shared decoder version. It is interesting to observe that in the latent space obtained by *CM-VAE*, the visual embedding degenerates into a purely non-informative Gaussian distribution as in [18]. *MS-VAE* alleviates the degeneration problem and learns similar distributions for audio and visual input, but the distance between two latent space is still evident. Wasserstein distance further bridges the two latent space and regularizes the learned data distributions. Note that the perfect alignment is very difficult to achieve and remains to be an open question in the research community.

4. CONCLUSION

In this paper, we propose the *MS-VAE* framework to learn the intrinsic audio-visual correlations for multiple downstream audio-visual tasks. *MS-VAE* is a self-supervised learning framework that generates a pair of corresponding audio-visual data given either one modality as input data. It leverages the advantage of label-free self-supervised learning from the generative models and achieves very competitive performance for multiple audio-visual tasks.

5. ACKNOWLEDGEMENTS

This research was partially supported by NSF NeTS-1909185, CSR-1908658 and Cisco. This article solely reflects the opinions and conclusions of its authors and not the funding agents. Yu Wu is supported by the Google PhD Fellowship.

6. REFERENCES

[1] Bruno Korbar, Du Tran, and Lorenzo Torresani, "Co-operative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018.

[2] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018.

[3] Relja Arandjelovic and Andrew Zisserman, "Objects that sound," in *ECCV*, 2018.

[4] Yapeng Tian, Chenxiao Guan, Goodman Justin, Marc Moore, and Chenliang Xu, "Audio-visual interpretable and controllable video captioning," in *CVPR Workshop*, 2019.

[5] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018.

[6] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba, "Self-supervised audio-visual co-segmentation," in *ICASSP*. IEEE, 2019.

[7] Di Hu, Feiping Nie, and Xuelong Li, "Deep multimodal clustering for unsupervised audiovisual learning," in *CVPR*, 2019.

[8] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba, "Music gesture for visual sound separation," in *CVPR*, 2020.

[9] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba, "Foley music: Learning to generate music from videos," in *ECCV*, 2020.

[10] Ruohan Gao, Rogerio Feris, and Kristen Grauman, "Learning to separate object sounds by watching unlabeled video," in *ECCV*, 2018.

[11] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *ECCV*, 2018.

[12] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NeurIPS*, 2016.

[13] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.

[14] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang, "Dual attention matching for audio-visual event localization," in *ICCV*, 2019.

[15] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu, "Deep cross-modal audio-visual generation," in *ACM Multimedia Workshop*, 2017.

[16] Aakanksha Rana, Cagri Ozcinar, and Aljosa Smolic, "Towards generating ambisonics using audio-visual cue for virtual reality," in *ICASSP*. IEEE, 2019.

[17] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[18] Huangjie Zheng, Jiangchao Yao, Ya Zhang, and Ivor W Tsang, "Degeneration in vae: in the light of fisher information loss," *arXiv preprint arXiv:1802.06677*, 2018.

[19] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *ICML*, 2017.

[20] Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush, "Latent alignment and variational attention," in *NeurIPS*, 2018.

[21] Yingtao Tian and Jesse Engel, "Latent translation: Crossing modalities by bridging generative models," *arXiv preprint arXiv:1902.08261*, 2019.

[22] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister, "Sliced and radon wasserstein barycenters of measures," *Journal of Mathematical Imaging and Vision*, vol. 51, no. 1, 2015.

[23] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf, "Wasserstein auto-encoders," in *ICLR*, 2017.

[24] 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 *ICASSP*, 2017.

[25] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang, "Dual-modality seq2seq network for audio-visual event localization," in *ICASSP*, 2019.

[26] Janani Ramaswamy and Sukhendu Das, "See the sound, hear the pixels," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020.

[27] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[28] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *ICCV*, 2017.

[29] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges, "Cross-modal deep variational hand pose estimation," in *CVPR*, 2018.