Learning Forgery-Aware Lip Representations Without Forgery Priors

Supplementary Material

A. Proof of Gaussian Distribution Matching

A.1. Preliminaries

Mutual Information (MI) is given by

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \quad (1)$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \quad (2)$$

or equivalently,

$$I(X;Y) = \mathrm{D}_{KL}\left(P_{X,Y} \parallel P_X \otimes P_Y\right), \quad (3)$$

where $H(X)$ is the entropy of X, quantifying its uncertainty; $H(X|Y)$ is the conditional entropy of X given Y, representing the residual uncertainty of X once Y is known; $H(X,Y)$ is the joint entropy of (X,Y). Eq. (3) is the exact Kullback Leibler(KL) divergence-based formulation of MI.

This formulation highlights that MI measures the reduction in entropy of one variable due to the knowledge of the other. When X and Y are independent, $I(X;Y) = 0$; otherwise, $I(X;Y) > 0$, with higher values indicating stronger dependency.

Maximize the MI between the inputs and outputs of a representation-learning function, i.e. an encoder, has been widely explored to learn viable low-dimensional representations of complex data [11]. In high-dimensional scenarios or when the true statistical distributions are unknown, directly computing entropies or their differences become infeasible. Fortunately, a number of estimation techniques have been developed to approximate MI. Among them, lower bounds are widely used to reparametrize the objective. In this work, we only consider a popular alternative based on the ***Donsker-Varadhan*** representation (DV, [7]) of the KL-divergence suggested in [1, 2]:

$$I(X;Y) = \sup_{T:\Omega \to R} \left[\mathrm{E}_{P_{X,Y}} T - \log \mathrm{E}_{P_X \otimes P_Y} \exp(T)\right] \quad (4)$$

where Ω is the sampling space, and T is a measurable critic function.

Deep InfoMax [11] suggests learning the most informative embeddings via a direct maximization of the mutual information between the original data and compressed representations:

$$I(X; f(X)) \to max, \quad (5)$$

where X is the random vector to be compressed, and f is the encoding mapping (being learned). In addition, data augmentation (crop randomly, add noise, etc.) are employed to inject stochasticity and avoid degenerate solutions. Consider the following Markov chain:

$$f(X) \to X \to X' \to f(X'), \quad (6)$$

with $I(f(X'); f(X)) \leq I(X'; f(X))$ being the new infomax objective.

To achieve Gaussian distribution matching of the latent space, we further modify Eq. (6) by adding independent noise Z to normalized representations of X, thus replacing $f(X)$ by $f(X) + Z$ in the infomax objective:

$$\begin{aligned} I(f(X); f(X)+Z) &= h(f(X)+Z) - h(f(X)+Z|f(X)) \\ &= h(f(X)+Z) - h(Z). \end{aligned} \quad (7)$$

with Z being independent of X.

In this case, maximizing the infomax objective is reduced to maximizing the entropy $h(f(X)+Z)$. According to the maximum entropy property of Gaussian distribution:

Theorem 1 (Theorem 8.6.5 in [4]). *Let X be a d-dimensional absolutely continuous random vector with probability density function p, mean m and covariance matrix Σ. Then*

$$h(X) = h\left(\mathcal{N}(m,\Sigma)\right) - \mathrm{D}_{KL}\left(p \parallel \mathcal{N}(m,\Sigma)\right),$$

$$h\left(\mathcal{N}(m,\Sigma)\right) = \frac{1}{2}\log\left((2\pi e)^d \det\Sigma\right),$$

where $\mathcal{N}(m,\Sigma)$ is a Gaussian distribution of mean m and covariance matrix Σ.

If we restrict $f(X)$ by fixing its covariance matrix, $h(f(X)+Z)$ attains its maximal value precisely when $f(X)+Z$ is distributed normally. As Z is independent to X, $f(X)+Z$ is normally distributed if and only if the distribution of $f(X)$ is also Gaussian. This forms a basis of the proposed method for Gaussian distribution matching.

A.2. Complete Proof of Proposition 1

Proposition 1 (Gaussian distribution matching). *Let $Z \sim \mathcal{N}(0, \sigma^2 I)$ and assume that each dimension of $f(X)$ is zero-mean with unit variance. Then the mutual information between $I(f(X); f(X)+Z)$ can be upper bounded as:*

$$I(f(X); f(X)+Z) \leq \frac{d}{2}\log\left(1 + \frac{1}{\sigma^2}\right), \quad (8)$$

with equality holding when $f(X) \sim \mathcal{N}(0, I)$.

Proof of Proposition 1.

Using Theorem1, we can rewrite Eq. (7), yielding

$$\begin{aligned} I(f(X); f(X)+Z) &= h\left(\mathcal{N}(m,\Sigma)\right) \\ &\quad - \mathrm{D}_{KL}\left(f(X)+Z \parallel \mathcal{N}(m,\Sigma)\right) - h(\mathcal{N}(0,\sigma^2 I)) \end{aligned}$$

where m and Σ are the mean and covariance matrix of $f(X)+Z$.

To bound the KL-divergence:

$$\begin{aligned}\mathrm{D}_{KL}\left(f(X)+Z \parallel \mathcal{N}(m,\Sigma)\right) = h\left(\mathcal{N}(m,\Sigma)\right) \\ - h(\mathcal{N}(0,\sigma^2 I)) - I(f(X); f(X)+Z)\end{aligned}$$

Next, we estimate the difference between the entropies by observing that

$$\begin{aligned}h\left(\mathcal{N}(m,\Sigma)\right) \leq \sum_{i=1}^{d} h\left(\mathcal{N}\left(m_i, \mathrm{Var}(f(X)_i) + \sigma^2\right)\right) \\ = d \cdot h\left(\mathcal{N}(0, 1+\sigma^2)\right)\end{aligned}$$

with the equality holding if and oly if Σ is diagonal, which implies $\Sigma = I$ since $\mathrm{Var}(f(X)_i) = 1$ for all $i \in \{1, ..., d\}$.

Therefore,

$$\begin{aligned}d \cdot h\left(\mathcal{N}(0,1+\sigma^2)\right) - h(\mathcal{N}(0,\sigma^2 I)) \\ = \frac{d}{2}\left[\log\left(1+\frac{1}{\sigma^2}\right) - \log\sigma^2\right] = \frac{d}{2}\log\left(1+\frac{1}{\sigma^2}\right)\end{aligned}$$

Finally,

$$\begin{aligned}\mathrm{D}_{KL}\left(f(X)+Z \parallel \mathcal{N}(m,\Sigma)\right) \leq \frac{d}{2}\log\left(1+\frac{1}{\sigma^2}\right) \\ - I(f(X); f(X)+Z)\end{aligned}$$

Since KL-divergence is non-negative, we obtain the desired bound:

$$I(f(X); f(X)+Z) \leq \frac{d}{2}\log\left(1+\frac{1}{\sigma^2}\right),$$

This concludes the proof.

B. More Implementation Details

B.1. Baselines

For the baselines we consider, we provide details on our implementations that are not given in the main text. Unless stated otherwise, the batch size is set to 32.

LipForensics/LipForen [9]. We use the official implementation[1]. Since the training scripts are not provided, we directly employ the pre-trained checkpoints and conduct inference under the same experimental settings.

SA-DTH [17]. As the official implementation is not publicly released, we re-implement the method following the paper and train it on the GRID dataset using supervised learning. For other datasets, we fine-tune the pre-trained model on the corresponding user data.

CIDE [8]. As the official implementation is unavailable, we re-implement the method following the paper and train it on the GRID dataset using multi-task learning. For other datasets, we fine-tune the pre-trained model on the corresponding user data.

Siamese [12]. Following the official implementation[2], we re-implement the model with PyTorch and conduct feature learning.

LipInc [5]. We use the official code[3] and employ the unified AV-HuBERT pretrained model as the visual encoder.

TD-VSA [10]. We use the official code[4] and train it on the VSA dataset [16] as recommended. We further fine-tune the identity branch to extract identity features for subjects in other datasets.

OpenSet [6]. As the official implementation is unavailable, we re-implement the method following the paper. We extract visual features using the unified AV-HuBERT pretrained model and train all user-specific models.

DO2HSC [19]. We use the official code[5] and employ the unified AV-HuBERT pretrained model as the visual encoder.

SpeechForensics/SpchForen [13]. We use the official inference code[6] and, while the original method also uses AV-HuBERT, we replace its checkpoint with the unified pretrained version used throughout our paper.

AVH-align [15]. We use the official code [7] and finetune the released checkpoint with the training data of the identities evaluated in our experiments.

B.2. Our Audio-Visual Variant

Following the implementation in [15], we concatenate the features from the visual and audio branches. Two projection layers are applied to map the original 1024-dimensional visual and audio features to 512 dimensions, which are then concatenated to form a 1024-dimensional joint representation. Since the final feature dimension matches our main model, the subsequent training strategy remains unchanged. In addition, because our method does not focus on audio feature augmentation, we simply mix the corresponding audio features using the same intensity coefficient as the visual features, without applying any additional operations.

C. Training Time Efficiency

Tab. 1 summarizes the average training time across all users involved in the AVLips dataset [14]. As shown, our approach achieves the most competitive performance while requiring noticeably less training time.

[1] https://github.com/ahaliassos/LipForensics

[2] https://github.com/deepconvolution/LipNet

[3] https://github.com/skrantidatta/LIPINC

[4] https://github.com/heyi0616/TDVSA

[5] https://github.com/wownice333/DOHSC-DO2HSC

[6] https://github.com/Eleven4AI/SpeechForensics

[7] https://github.com/bit-ml/AVH-Align

Table 1. Evaluation of training time efficiency (Time (s)↓) on TFG-GRID.

Method	Time / Epoch (s)	#Epochs	Total Time (s)
SA-DTH [17]	200	28	5,600
CIDE [8]	1,320	11	14,520
Siamese [12]	25	21	525
LipInc [5]	160	29	4,640
TD-VSA [10]	820	30	24,600
OpenSet [6]	48	38	1,824
DO2HSC [19]	6.5	250	1,625
AVH-align [15]	158	33	5,214
Ours	18	45	810

D. Code and Proposed TFG-Suite

Code and dataset: https://github.com/blair00822/VSA.

E. More Ablations

We provide comprehensive ablations on pseudo-forgery generation strategies in Tab. 2, which consistently validate the effectiveness of our design. All experiments are conducted using 50 training samples per user to ensure a consistent evaluation protocol.

Table 2. Ablations (AUC (%)↑) of pseudo-forgeries on AVLips.

Method	CutMix [18]	ProDet [3]	w/ SML	w/ CML	**Ours**
HM	90.48	90.99	95.00	94.79	**95.56**
fake	93.12	94.07	95.61	96.59	**98.32**

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, et al. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 531–540. PMLR, 2018. 1

[2] Ivan Butakov, Alexander Semenenko, Alexander Tolmachev, et al. Efficient distribution matching of representations via noise-injected deep infomax. In *ICLR*, 2025. 1

[3] Jikang Cheng, Zhiyuan Yan, Ying Zhang, et al. Can we leave deepfake data behind in training deepfake detector? *NeurIPS*, 37:21979–21998, 2024. 3

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006. 1

[5] Soumyya Kanti Datta, Shan Jia, and Siwei Lyu. Exposing lip-syncing deepfakes from mouth inconsistencies. In *ICME*, pages 1–6. IEEE, 2024. 2, 3

[6] Michael Macedo Diniz and Anderson Rocha. Open-set deepfake detection to fight the unknown. In *ICASSP*, pages 13091–13095. IEEE, 2024. 2, 3

[7] Monroe D Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1): 1–47, 1975. 1

[8] Zihao Guo and Shilin Wang. Content-insensitive dynamic lip feature extraction for visual speaker authentication against deepfake attacks. In *ICASSP*, pages 1–5, 2023. 2, 3

[9] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, et al. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049, 2021. 2

[10] Yi He, Lei Yang, Shilin Wang, et al. Lip feature disentanglement for visual speaker authentication in natural scenes. *IEEE TCSVT*, 34(10):9898–9909, 2024. 2, 3

[11] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, et al. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 1

[12] Brando Koch and Ratko Grbić. One-shot lip-based biometric authentication: Extending behavioral features with authentication phrase information. *Image and Vision Computing*, 142: 104900, 2024. 2, 3

[13] Yachao Liang, Min Yu, Gang Li, et al. Speechforensics: Audio-visual speech representation learning for face forgery detection. *NeurIPS*, 37:86124–86144, 2024. 2

[14] Weifeng Liu, Tianyi She, Jiawei Liu, et al. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. *NeurIPS*, 37:91131–91155, 2024. 2

[15] Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, et al. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *CVPR*, pages 18815–18825, 2025. 2, 3

[16] Jiahui Sun, Shilin Wang, and Quanhai Zhang. Visual speaker authentication by a cnn-based scheme with discriminative segment analysis. In *Neural Information Processing*, pages 159–167. Springer International Publishing, 2019. 2

[17] Chen-Zhao Yang, Jun Ma, Shilin Wang, et al. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, 16:1841–1854, 2020. 2, 3

[18] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*, pages 6023–6032, 2019. 3

[19] Yunhe Zhang, Yan Sun, Jinyu Cai, et al. Deep orthogonal hypersphere compression for anomaly detection. *arXiv preprint arXiv:2302.06430*, 2023. 2, 3