UIA-ViT: Unsupervised Inconsistency-Aware Method based on Vision Transformer for Face Forgery Detection Supplementary Material

Wanyi Zhuang, Qi Chu*, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu

CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China
wy970824@mail.ustc.edu.cn, qchu@ustc.edu.cn, {tzt,liuqk3,doubihj, miaoct,zxluo}@mail.ustc.edu.cn, ynh@ustc.edu.cn.

1 Method

Unsupervised Approximate Forgery Location. During training, F_{real} and F_{fake} are updated by new (μ_r,Σ_r) and (μ_f,Σ_f), which are approximated with the sample mean and sample covariance from the observations ($x_r^1, x_r^2, ..., x_r^n \in \mathbb{R}^D$) and ($x_f^1, x_f^2, ..., x_f^n \in \mathbb{R}^D$). We accumulate the feature observations from every mini batch of training samples and experimentally update two MVG distributions every 0.5 training epochs :

$$\mu_r = \frac{1}{n}\sum_{i=1}^{n} x_r^i, \quad \Sigma_r = \frac{1}{n-1}\sum_{i=1}^{n}(x_r^i - \mu_r)(x_r^i - \mu_r)^T; \tag{1}$$

$$\mu_f = \frac{1}{n}\sum_{i=1}^{n} x_f^i, \quad \Sigma_f = \frac{1}{n-1}\sum_{i=1}^{n}(x_f^i - \mu_f)(x_f^i - \mu_f)^T. \tag{2}$$

2 Experiment

2.1 Determine Which Layer for Patch Consistency Learning

We conduct experiments on this issue: the Attention Map of which layers are proper to be confined to consistency-related pattern. As shown in Table.1, several experiments are conducted in the way to utilize the mean Attention Map from single 8-th, 9-th, 10-th, 11-th, 12-th layer for patch consistency learning in UPCL module and as consistency weighted matrix in PCWA module. We find that it is effective to restrict every single layer comparing to baseline. To further retain useful information from different layers as much as possible, we use the mean Attention Map from layer 8 to 12 for UPCL and PCWA modules, which

* Corresponding authors.

Table 1. Experiments on different layers to conduct UPCL loss

	Layer	-	8	9	10	11	12	all 8-12
FF++	ACC(%)	95.86	97.43	96.86	97.14	97.14	97.29	**97.43**
	AUC(%)	99.30	99.21	99.19	99.25	99.07	99.11	**99.33**
Celeb-DF-v2	AUC(%)	76.25	80.71	79.26	80.22	80.78	80.33	**82.41**

Table 2. Performance comparison with different UPCL loss weight λ_1

	λ_1	0.03	0.04	0.05	**0.06**	0.07	0.08	0.09
FaceForensics++	ACC(%)	96.86	97.00	97.14	**97.43**	96.71	97.14	96.71
	AUC(%)	99.20	99.24	99.28	**99.33**	99.17	99.33	99.23
Celeb-DF-v2	AUC(%)	80.65	79.43	80.60	**82.41**	80.85	80.39	80.65

achieve better performance both on intra-dataset (FF++) detection and cross-dataset (Celeb-DF-v2) detection. It demonstrates the conjunction with multiple attention maps from different layers is benefit for network to capture consistency-related pattern.

2.2 Loss Weights

The total loss functions of the proposed method are described as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{UPCL} + \lambda_2 (\frac{1}{|c1|} + \frac{1}{|c2|}) + \lambda_3 |c3|. \tag{3}$$

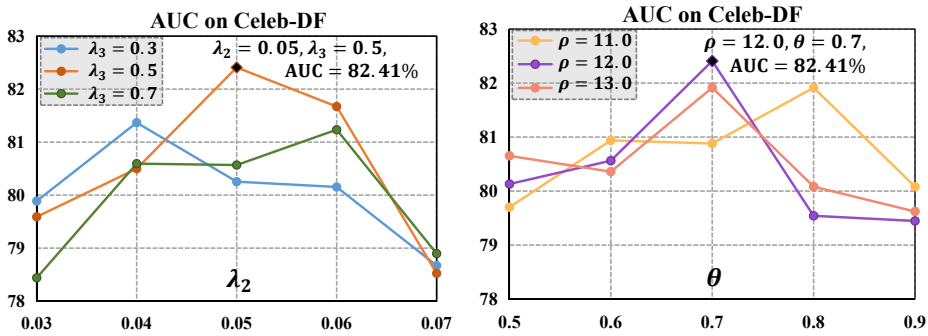
We explore the appropriate weights for UPCL loss in Eqn. (3). As shown in Table.2, we set a series of UPCL loss weights λ_1, then compare their performance on the FaceForensics++(high quality) detection task and generalization ability on Celeb-DF. The performance shows a trend of rising first and then falling with the gradient λ_1 and the best detection accuracy is achieved when λ_1 is set as 0.06.

Moreover, we explore the loss weights λ_2, λ_3 for three learnable parameters c_1, c_2, c_3 in Eqn. (3). c_1, c_2, c_3 are designed to generate soft pseudo label for mean Attention Map Υ^P based on predicted location map $\mathbf{M}$, formalized as:

$$\mathbf{C}_{(i,j),(k,l)} = \begin{cases} c_1, if\ \mathbf{M}_{ij} = 0\ and\ \mathbf{M}_{kl} = 0 \\ c_2, if\ \mathbf{M}_{ij} = 1\ and\ \mathbf{M}_{kl} = 1 \\ c_3, \qquad else \end{cases}. \tag{4}$$

In the total loss Eqn.(3), the second last item is designed for optimizing c_1, c_2 to increase and the last item is designed for optimizing c_3 to decrease along the training stage. Although c_1, c_2, c_3 don't be directly restricted within a certain range , quite amounts of experiments show each of them wouldn't go out of range $[0, 1]$.

We explore the proper weights λ_2, λ_3 to control the convergence rate of three learnable parameters c_1, c_2, c_3, where c_1, c_2 share the same loss weight λ_2. As



(a) Loss Weights (b) Progressive Weighted Function

Fig. 1. Parameter Experiments.

shown in Fig.1(a), we set a series of loss weights λ_2, λ_3, and train all models on FaceForensics++(high quality). Comparing their AUC performance on the testing set Celeb-DF-v2, the best detection AUC is achieved when λ_2 and λ_3 are set as 0.05 and 0.5. And after convergence, c_1, c_2 and c_3 eventually tend to be near $(0.8, 0.8, 0.0)$.

2.3 Progressive Weighted Function of PCWA

In PCWA module, the weighted matrix $\mathbf{A}^P$ gradually transfers from averaged weighting (all-one matrix) to consistency weighting. We design the progressively decreasing function in the range of $(0, 1)$ with hyper-parameters ρ and θ for variable weight w, as shown in Eqn.(5).

$$w = \text{sigmoid}(-\rho(step - \theta)), step = \frac{current\_iters}{total\_iters} \in [0, 1], \tag{5}$$

$$\mathbf{A}^P = w * \mathbb{1} + (1 - w) * \text{sigmoid}(\Upsilon^C). \tag{6}$$

We explore the suitable ρ and θ to control the transition speed of weighted matrix $\mathbf{A}^P$. As shown in Fig.1(b), we set a series of ρ and θ, and also train on FaceForensics++(high quality). Comparing their AUC performance on Celeb-DF-v2, we find the best result when ρ and θ are set as 12.0 and 0.7.