Exploring Disentangled Content Information for Face Forgery Detection

Jiahao Liang Huafeng Shi Weihong Deng*

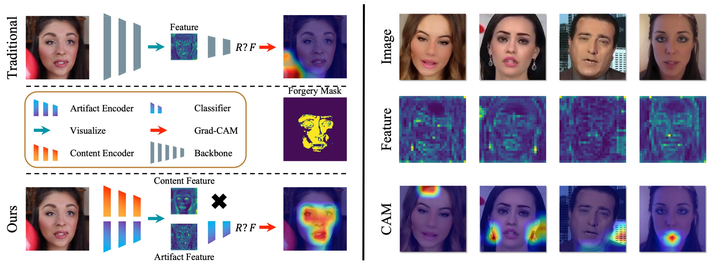
Beijing University of Posts and Telecommunications SenseTime Research

Motivation

- The detector is prone to focus more on content information than artifact traces, suggesting that the detector is sensitive to the intrinsic bias of the dataset, which leads to severe overfitting.

Basic Idea

- Disentangle content features and artifact features, and only the disentangled artifact features for face forgery detection, thus ignoring the interference of content information.



Comparison with ordinary methods

Visualization of ordinary methods

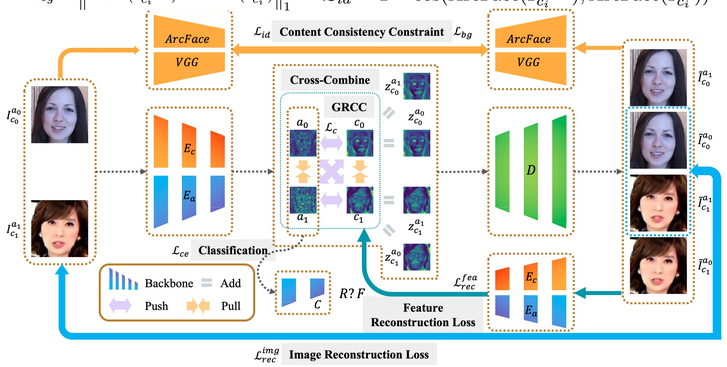
Main Contributions

- Explore the impact of content information on the generalization performance of face forgery detection.
- Design an easily embeddable disentanglement framework for content information removal.
- Propose a Content Consistency Constraint (C2C) and Global Representation Contrastive Constraint (GRCC) to enhance the independence of disentangled features.

Proposed Disentanglement Framework

1. Ensure the consistency of content information:

$$\mathcal{L}_{bg} = \left\| \mathrm{VGG}(\widetilde{I}^{a_{1-i}}_{c_i}) - \mathrm{VGG}(I^{a_i}_{c_i}) \right\|_1 \qquad \mathcal{L}_{id} = 1 - \cos(\mathrm{ArcFace}(\widetilde{I}^{a_{1-i}}_{c_i}), \mathrm{ArcFace}(I^{a_i}_{c_i}))$$



2. Ensure the consistency at pixel level and feature level:

$$\mathcal{L}^{img}_{rec} = \sum_{i=0}^{1} \left\| I^{a_i}_{c_i} - \widetilde{I}^{a_i}_{c_i} \right\|_1 \quad \mathcal{L}^{fea}_{rec} = \sum_{i=0}^{1} \left(\left\| E_c(\widetilde{I}^{a_i}_{c_i}) - c_i \right\|_1 + \left\| E_a(\widetilde{I}^{a_i}_{c_i}) - a_i) \right\|_1 + \left\| E_c(\widetilde{I}^{a_i}_{c_{1-i}}) - c_{1-i} \right\|_1 + \left\| E_a(\widetilde{I}^{a_i}_{c_{1-i}}) - a_i) \right\|_1$$

3. Take the advantage of the InfoNCE to construct a Global Representation Contrastive Constraint (GRCC) between the artifact and content features:

$$\mathcal{L}_c = -\log[\frac{\exp(d(\mathbf{G}_{a_0}, \mathbf{G}_{a_1}))}{\exp(d(\mathbf{G}_{a_0}, \mathbf{G}_{a_1})) + \sum_{i=0}^{1} \exp(d(\mathbf{G}_{a_i}, \mathbf{G}_{c_{1-i}}))}] - \log[\frac{\exp(d(\mathbf{G}_{c_0}, \mathbf{G}_{c_1}))}{\exp(d(\mathbf{G}_{c_0}, \mathbf{G}_{c_1})) + \sum_{i=0}^{1} \exp(d(\mathbf{G}_{a_i}, \mathbf{G}_{c_{1-i}}))}]$$

Overall Loss: $\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}^{img}_{rec} + \lambda_2 \mathcal{L}^{fea}_{rec} + \lambda_3 \mathcal{L}_{id} + \lambda_4 \mathcal{L}_{bg} + \lambda_5 \mathcal{L}_c$

Experiments

In-Dataset evaluation (ACC (%)) on FF++ (LQ). We combine each forgery and real dataset in pairs to construct four sub-datasets, and evaluate the corresponding performance. After embedding into our framework, all detectors achieve considerable performance gains and even outperform other methods. AVG: the average performance of the four sub-datasets.

Method	DF	FF	FS	NT	AVG
Steg.Features [16]	67.00	48.00	49.00	56.00	55.00
LD-CNN [12]	75.00	56.00	51.00	62.00	61.00
C-Conv [5]	87.00	82.00	74.00	74.00	79.25
CP-CNN [37]	80.00	62.00	59.00	59.00	65.00
MesoNet [3]	90.00	83.00	83.00	75.00	82.75
F^3-Net [36]	96.81	94.01	95.85	79.36	91.51
Gram-Net [27]	95.12	88.01	93.34	76.12	88.15
+ Ours	**95.67**	**89.06**	**94.01**	**76.96**	**88.93**
RFM [44]	95.42	91.24	93.60	79.83	90.02
+ Ours	**95.92**	**92.27**	**93.97**	**80.14**	**90.58**
ResNet-50 [18]	95.23	87.79	92.34	76.28	87.91
+ Ours	**95.43**	**88.94**	**93.99**	**77.19**	**88.89**
Xception [9]	95.36	91.94	93.55	78.32	89.79
+ Ours	**96.50**	**93.62**	**94.76**	**79.02**	**90.98**
ResNest-50 [49]	95.98	92.16	93.13	78.22	89.87
+ Ours	**98.95**	**94.32**	**94.56**	**80.46**	**92.10**

Cross-Dataset evaluation on Celeb-DF (AUC (%)) by training on FF++- DF (ACC (%)). Our method outperforms all the methods with the same backbone (Xception) and achieves the best performance with the backbone of ResNest-50.

BackBone	Method	FF++-DF (Train)	Celeb-DF (Test)
Xception	F^3-Net [36]	97.97	65.17
Efficient-B4	Zhao *et al.* [53]	-	67.44
HRNet	Face X-ray [31]	-	74.76
Xception	SPSL [25]	96.91	76.88
-	Chen *et al.* [8]	98.84	78.26
ResNet-18	Gram-Net [27]	95.12	67.14
	+ Ours	**95.67**	**74.94**
Xception	RFM [44]	95.42	67.21
	+ Ours	**95.92**	**74.44**
ResNet-50	ResNet-50 [18]	95.23	66.84
	+ Ours	**95.43**	**74.71**
Xception	Xception [9]	95.36	65.50
	+ Ours	**96.50**	**76.91**
ResNest-50	ResNest-50 [49]	95.98	68.00
	+ Ours	**98.95**	**82.38**

Comparison of our framework with baseline methods on the identity and background unbalanced dataset.

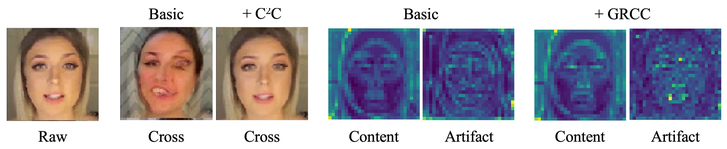
Method	ID Unbalanced Dataset				BG Unbalanced Dataset			
	ACC	AUC	EER	$TDR_{0.1}$	ACC	AUC	EER	$TDR_{0.1}$
Gram-Net [27]	89.85	96.49	10.14	89.70	79.84	87.83	20.19	66.10
+ Ours	**94.80**	**99.48**	**3.619**	**98.90**	**94.00**	**98.93**	**5.764**	**96.70**
RFM [44]	90.34	95.34	9.232	90.93	85.09	92.33	14.89	79.58
+ Ours	**95.49**	**99.11**	**4.102**	**97.90**	**95.02**	**98.34**	**4.839**	**96.93**
ResNet-50 [18]	89.61	96.46	10.35	89.30	80.88	91.29	17.17	72.30
+ Ours	**95.39**	**99.54**	**3.571**	**99.00**	**94.46**	**98.76**	**5.524**	**97.20**
Xception [9]	91.06	96.91	8.967	91.50	84.39	92.42	17.17	78.10
+ Ours	**95.85**	**99.32**	**3.434**	**99.10**	**95.14**	**98.71**	**4.762**	**97.00**
ResNest-50 [49]	89.89	97.54	8.507	92.70	81.28	93.68	14.34	79.60
+ Ours	**95.58**	**99.61**	**3.190**	**98.90**	**94.56**	**98.48**	**5.479**	**96.90**

- We conduct comparative experiments on these datasets, and the experimental results are shown in Table. We can find that the performance on the ID and BG unbalanced datasets suffers a huge drop, which indicates that the existence of the intrinsic bias does interfere with the optimization of the detector. In contrast, our framework can maintain a high performance even on the unbalanced dataset by stripping the content features and thus eliminating the interference of content bias. Furthermore, compared with the ID unbalanced dataset, the performance degradation on the BG unbalanced dataset is more serious.

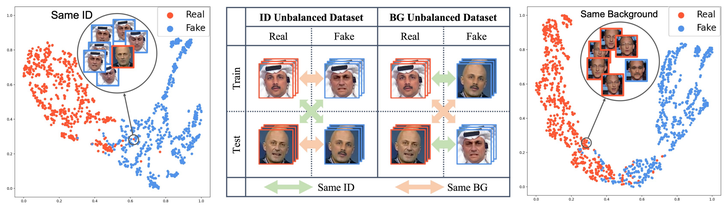
Ablation study on the FF++-DF and Celeb-DF. "Basic" represents the basic disentanglement framework.

Method	Basic	C^2C	GRCC	FF++-DF	Celeb-DF
Xception				95.36	65.50
Variant A	✓			94.55	65.11
Variant B	✓	✓		95.73	70.08
Variant C	✓		✓	96.33	72.57
Variant D	✓	✓	✓	**96.50**	**76.91**

Visualizations

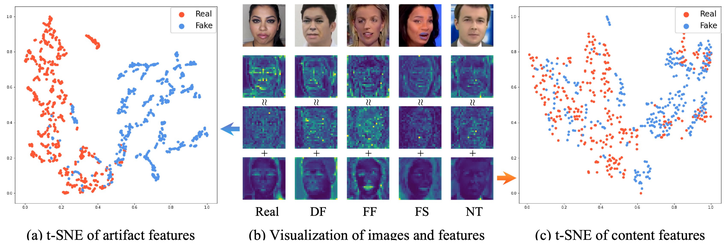


Visualization of the ablation study, which illustrates the impact of C2C on the reconstructed images and GRCC on the disentangled features. "Raw" represents the raw image, and "Cross" represents the cross-reconstruction image.



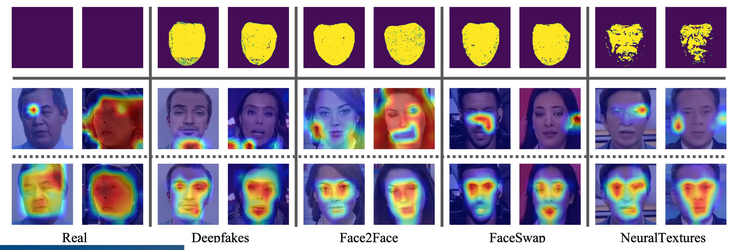
(a) t-SNE of ID unbalanced dataset (b) Construction process of unbalanced datasets (c) t-SNE of BG unbalanced dataset

(b) **The construction process** of unbalanced datasets. (a)(c) **t-SNE feature visualization** of the Xception network on the ID and BG unbalanced dataset.



(a) t-SNE of artifact features (b) Visualization of images and features (c) t-SNE of content features

(b) **Visualization** of the image (first row), traditional detector's (Xception) features (second row), ours disentangled artifact (third row) and content features (fourth row). (a)(c) **t-SNE visualization** of artifact features and content features.



Visualization of forgery mask (first row), Xception's (second row) and ours (third row) Grad-CAM on five sub-datasets of FF++. Grad-CAM shows that the baseline is prone to overfitting to small local regions or focusing on content noise outside the forgery region. In contrast, the activation region of our method is comprehensive and almost consistent with the forgery mask.