

Supplementary Materials: Boosting Semantic Disentanglement: An Unconventional Harnessing of LVM for Open World Deepfake Interpretation

Anonymous Author(s)
Submission Id: 1702

In this supplementary material, we provide,
1. More details of OW-DFI in Section 1.
2. Details of comparison methods in Section 2.
3. Visualization of our proposed DIN in Section 3.
4. Conclusion in Section 4.

1 DETAILS OF OW-DFI

To address the evolving challenges posed by Artificial Intelligence Generated Content (AIGC) in open-world settings, we introduce a new task with corresponding benchmarks, named Open World Deepfake Interpretation (OW-DFI). The objective of this task is to enhance the interpretation capabilities for evolving deepfakes by providing extensive benchmarks. The OW-DFI benchmarks comprise 11,499 genuine images and 14,319 artificial images, sourced from 24 distinct AIGC methods across four disparate datasets. In Figure 1, we provide a visualization of the diverse range of manipulated objects. These include human features such as hair, face, eyes, mouth, nose, eyebrows, and eyeglasses, as well as background elements like bedrooms, churches, natural scenes, and animals. This visualization emphasizes the extensive nature of manipulable semantics, underscoring the necessity for precise localization and addressing complex semantic interference in forgery detection. Moreover, Figure 2 shows the application of different forgery techniques to the same object. This example demonstrates the advanced capabilities of AIGCs to execute intricate forgeries, emphasizing the increasing necessity for interpretation models that can not only detect but also localize and attribute forgeries effectively.

2 DETAILS OF COMPARISON METHODS

To assess the effectiveness of our proposed Deepfake Interpretation Network (DIN) within the OW-DFI framework, we compare it against three state-of-the-art (SOTA) methods: LVNet [3], POP [1], and OW-DFA [4]. The comparisons encompass a spectrum of approaches, ranging from the conventional deepfake interpretation technique to those that are specifically tailored for open-world scenarios.

LVNet. LVNet [3] is designed to localize and verify artifacts by integrating RGB and SRM noise modalities through an advanced fusion process. The architecture of LVNet comprises a feature encoder for modality fusion, complemented by distinct localization and classification modules to facilitate comprehensive interpretation. Originally, as the classification module of LVNet is configured for binary classification, we extend the classifier to support attribution across 25 distinct classes. During the evaluation process, attributions are first executed based on the K known classification weights, and then these results are combined with binary localization outcomes to produce pixel-level interpretations. For unknown

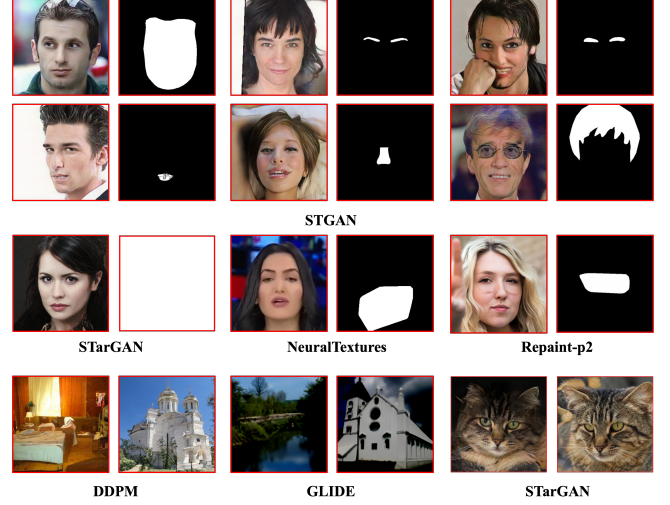


Figure 1: The visualization of the manipulated images with corresponding masks, encompasses a diverse range of manipulable objects. The manipulated regions are indicated by white pixels in even columns and are maintained across subsequent figures.

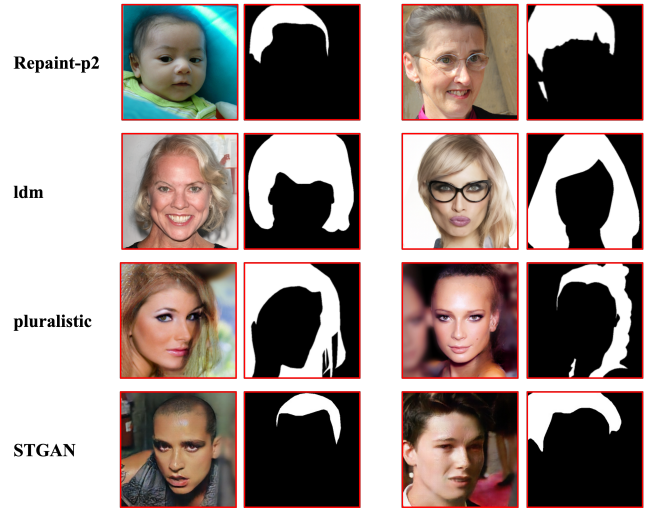


Figure 2: The visualization of images manipulated by various forgery techniques.

forgery detection, we calculate the unknown confidence score using the Maximum Softmax Probability (MSP). This score is integrated with the binary localization results to produce pixel-level interpretations. LVNet was initially trained on the base set with a batch size of 16 and a learning rate of $1e-4$ using the Adam optimizer. Training continued until no further reduction in loss was observed within three epochs. During the incremental learning phase, with a limited number of training classes, we fine-tuned the classification weights over 50 epochs.

POP. POP[1] aims to learn orthogonal prototypes to discern regions associated with novel classes without affecting old classes in an open-world environment. To localize each known class, the relevant prototype is projected onto the entire feature map to generate class-specialized features, which are then processed by a shared binary classifier. During the base training phase, POP was trained with a batch size of 16 for 50 epochs at a learning rate of $1e-3$, utilizing the SGD optimizer. In the incremental learning phase, the system was specifically focused on adapting novel class prototypes to be orthogonal to existing prototypes, extending over an additional 50 epochs.

OW-DFA. OW-DFA[4] is designed to identify unseen categories within mixed datasets that include both labeled and unlabeled data. The training protocol employs 7 base categories as labeled data, while the data from the remaining 17 categories is treated as unlabeled. This model is trained over 50 epochs with a batch size of 128, using the Adam optimizer. The initial learning rate is set at $2e-4$ and is systematically reduced by 20% every 10 steps in accordance with the StepLR schedule.

The final comparative results highlight the superiority of our proposed DIN, demonstrating a significant reduction in the False Positive Rate (FPR)—by at least 35.05%—in localizing unknown forgeries. Additionally, there is at least a 5.25% improvement in the Intersection over Union (IOU) for localizing novel known categories after the third incremental session. These outcomes affirm the effectiveness of our approach in eliminating non-causal semantics by leveraging purified semantic priors from Large Visual Models (LVMs).

3 VISUALIZATION OF DIN

t-SNE Visualization of open-world deepfake interpretation.

To enable a more intuitive comparison of our Deepfake Interpretation Network (DIN), we conducted t-SNE visualization [2], contrasting it with OW-DFA [4], as illustrated in Figure 3. Our observations reveal that DIN creates a more distinct feature space, enabling it to more effectively differentiate natural images (marked in black) from forgeries (represented in various colors). This enhanced separation is advantageous for analyzing inter-forgery relationships, which is crucial for subsequent few-shot incremental learning.

t-SNE Visualization of few-shot incremental learning. To provide a more intuitive evaluation of our Deepfake Interpretation Network (DIN) under the evolving AIGCs, we utilize t-SNE visualization [2] to illustrate the network’s performance before and after few-shot incremental learning (FIL). These stages are depicted in Figure 4, with novel categories highlighted by red circles. The visualization clearly shows that DIN can effectively discern novel features

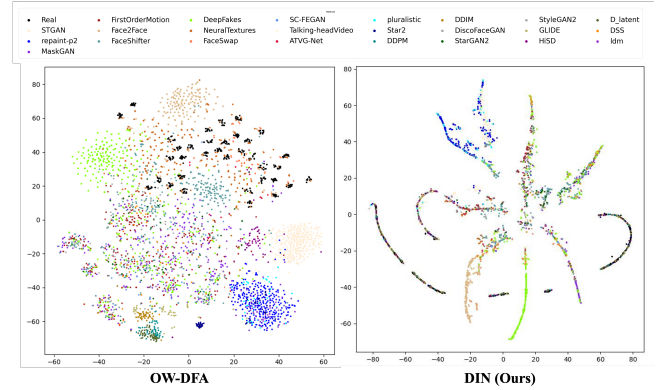


Figure 3: The t-SNE Visualization of feature distribution. This figure illustrates the feature space learned from 24 forgery categories after training on 7 forgery methods. For this visualization, 500 samples from the 7 base categories and 50 samples from unseen forgeries within the test dataset were randomly selected. This sampling configuration is consistently maintained across subsequent figures.

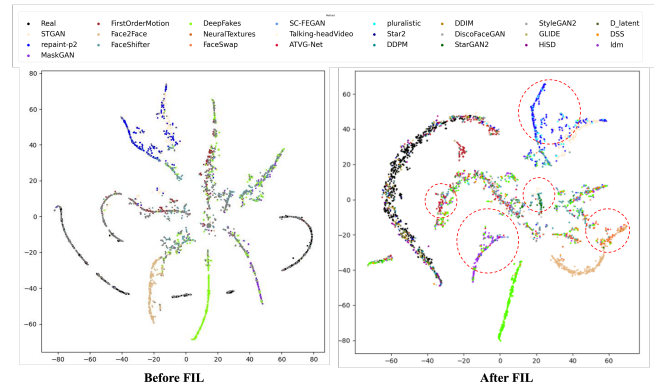


Figure 4: The t-SNE Visualization of Feature Distribution before and after Few-Shot Incremental Learning (FIL). Before FIL, the 7 base categories are represented in various colors, while the unknown categories are depicted in gray, as shown in the left panel of the figure. After FIL, known categories are distinguished by different colors and encircled in red, as illustrated in the right panel.

from just 5-shot training samples in an incremental learning setting. This result confirms the efficacy of the Correlation-based Incremental Module (CIM) in leveraging inherent inter-forgery correlations and the Semantic-prior Orientation Module (SOM) in mitigating non-causal semantics.

Visualization of interpretation results with diversely manipulated objects. We present visualization results of our DIN applied to images manipulated by STGAN, encompassing five types of manipulations: hair, nose, mouth, face, and eyes. The corresponding results, depicted in Figure 5, illustrate DIN’s capability to accurately localize the forged regions and attribute the specific AIGC technique. These outcomes underscore the significance of eliminating

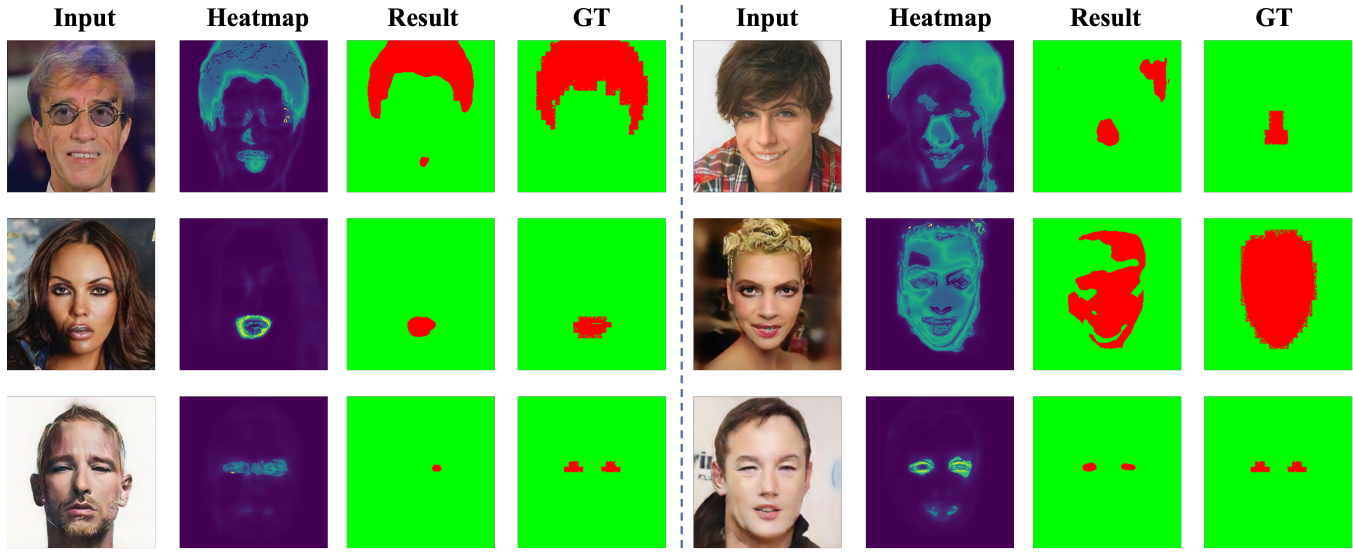


Figure 5: Visualization of interpretation results for forgeries with diversity manipulated objects.

non-relevant semantic factors to enhance the accuracy of forgery interpretations.

4 CONCLUSION

In this paper, we propose a novel task, Open-World Deepfake Interpretation (OW-DFI), designed to enhance the interpretability of deepfakes in dynamic open-world scenarios. It encompasses a wider variety of manipulated objects beyond just faces, addressing the complexities introduced by advanced forgery techniques. Our Deepfake Interpretation Network (DIN) leverages semantic insights from LVMs to mitigate reliance on spurious semantic patterns in a non-parametric manner. Experimental results validate the SOTA performance of our method across 24 forgery methods, underscoring its capability to generate reliable interpretations and uncover novel forgeries as forgery techniques evolve.

REFERENCES

- [1] Sun'ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. 2023. Learning Orthogonal Prototypes for Generalized Few-Shot Semantic Segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 11319–11328.
- [2] Pavlin Gregor Polcar, Martin Straar, and Bla Zupan. 2019. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *bioRxiv* (2019).
- [3] Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. 2023. Locate and verify: A two-stream network for improved deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7131–7142.
- [4] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. 2023. Contrastive Pseudo Learning for Open-World DeepFake Attribution. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), 20825–20835.