УДК 004.93'1

Мельников Виталий Андреевич, доцент кафедры информационных технологий и экономической информатики, кандидат экономических наук, Челябинский государственный университет, г. Челябинск

Перина Дарья Андреевна, магистрант, Челябинский Государственный Университет, г. Челябинск

ОБНАРУЖЕНИЕ ДИПФЕЙКОВ НА ОСНОВЕ ПРИЗНАКОВ DINO_V2 и MOBILE_NET_V4: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ФОТО- И ВИДЕОКЛАССИФИКАЦИИ

Аннотация

В работе исследуется задача автоматического обнаружения синтетических (deepfake) лиц на фотографиях и видеозаписях с применением двух семейств нейронных сетей — самонаблюдаемого трансформера DINOv2 и лёгкой сверточной сети MobileNetV4. Обучение проводилось на объединенном наборе данных из более 37 000 видеозаписей, включающем публичные бенчмарки FaceForensics++, Celeb-DF, DFDC, DeepFaceDetection, а также собственные материалы, синтезированные алгоритмами замены лиц (inswapper) и диффузионными моделями (Stable Diffusion XL). Предлагается двухуровневый подход: фреймовый классификатор распознаёт поддельные лица на отдельных изображениях, а видеоклассификатор агрегирует признаки по 16 кадрам через временно́е усреднение. Лучшая модель на основе DINOv2 достигает ROC-AUC = 0.9988 на тестовой выборке видеозаписей при общей точности 98.30%, превосходя аналог на MobileNetV4 на 1.66 п.п. При этом MobileNetV4 демонстрирует оптимальное соотношение между производительностью и

качеством. На уровне отдельных кадров фотоклассификатор DINOv2 демонстрирует полноту 99.51% для класса «реальные» и 99.73% для класса «поддельные».

Abstract

This paper investigates the problem of automated detection of synthetic (deepfake) faces in photographs and video recordings using two families of neural networks: the self-supervised transformer DINOv2 and the lightweight convolutional network MobileNetV4. Training was conducted on a combined dataset of over 37,000 video recordings, including public benchmarks such as FaceForensics++, Celeb-DF, DFDC, and DeepFaceDetection, as well as proprietary materials synthesized using face-swapping algorithms (inswapper) and diffusion models (Stable Diffusion XL). A two-level approach is proposed: a frame-based classifier identifies fake faces in individual images, while a video classifier aggregates features across 16 frames via temporal averaging. MobileNetV4 achieves 96.64% accuracy, demonstrating a favorable trade-off between efficiency and quality. The best-performing model, based on DINOv2, achieves a ROC-AUC of 0.9988 on the video test set with an overall accuracy of 98.30%, outperforming the MobileNetV4 counterpart by 1.66 percentage points. At the individual frame level, the DINOv2 photo classifier demonstrates a recall of 99.51% for the «real» class and 99.73% for the «fake» class.

Ключевые слова: обнаружение дипфейков, DINOv2, MobileNetV4, самонаблюдаемое обучение, видеоклассификация, визуальный трансформер.

Keywords: deepfake detection, DINOv2, MobileNetV4, self-supervised learning, video classification, vision transformer (ViT).

1. Введение

Технологии синтеза лиц на основе глубокого обучения (deepfake) развиваются стремительно: современные диффузионные модели и алгоритмы замены лиц позволяют создавать фотореалистичные видеозаписи с подменённым лицом человека. Подобный контент несёт серьёзные угрозы: от дискредитации публичных лиц и распространения дезинформации до мошенничества с использованием биометрической аутентификации [1, 2].

Задача обнаружения дипфейков формулируется как бинарная классификация: по входному изображению или видеозаписи необходимо определить, является ли лицо реальным или синтетическим. Сложность задачи обусловлена постоянным совершенствованием методов генерации, разнообразием условий съёмки, степенью компрессии видео и множеством алгоритмов подделки [10].

В данной работе предлагается двухуровневая система обнаружения дипфейков: фреймовый уровень (классификация отдельных кадров-лиц) и видеоуровень (агрегирование признаков по последовательности кадров). Для обоих уровней сравниваются трансформерный бэкбон DINOv2 ViT-S/14 и сверточный MobileNetV4-Conv-Medium. Модели обучаются на гетерогенном датасете, включающем публичные бенчмарки и оригинальные данные, синтезированные современными методами.

2. Обзор предшествующих работ.

Ранние подходы к обнаружению дипфейков опирались на явные артефакты — мерцание по краям лица, неестественное моргание, несогласованность текстуры [3]. Переход к нейросетевым методам показал, что свёрточные сети (VGG, Xception) хорошо обобщаются внутри одного метода синтеза, но теряют качество при переносе на неизвестные атаки [5, 6]. Трансформерные детекторы на базе ViT [11] фиксируют глобальные зависимости между патчами лица, недоступные локальным свёрткам, а самонаблюдаемое предобучение DINOv2 [6] позволяет получать мощные

признаковые экстракторы без разметки — что особенно ценно при ограниченных объёмах данных о дипфейках. На видеоуровне применяются агрегирование признаков через LSTM, Transformer-энкодеры или среднее усреднение [9, 10]; настоящая работа исследует последний подход в сочетании с современными бэкбонами.

3. Набор данных

3.1 Источники данных

Обучение проводилось на объединенном датасете (табл. 1). Публичная часть включает FaceForensics++ (F++) [8] с четырьмя методами манипуляции (Deepfakes, FaceSwap, Face2Face, FaceShifter), Celeb-DF [4] с высококачественными дипфейками знаменитостей, DFDC [1] с разнообразными условиями съёмки и DeepFaceDetection (DFD). Оригинальная часть включает видео, синтезированные алгоритмом InsightFace inswapper_128 (Dataset_Dasha_inswapper, Dataset_Real_1_inswapper), а также с помощью Stable Diffusion XL + IP-Adapter (Dataset_Dasha1_sdxl, Dataset_Real_0_sdxl), и оригинальные записи реальных лиц (Real_new_video, Real_Dasha, Real_Dasha1).

Таблица 1. Распределение данных по разбивкам

Разбивка	Всего
Train	32595
Val	5750
Test	5417
Итого	43762

3.2 Предобработка

Для видеоданных применялась следующая процедура: детекция лиц с помощью MediaPipe, вычисление медианного bounding box по всем кадрам видео, обрезка с отступом 40%, фильтрация некачественных кадров (чёрные, пересвеченные, однородные) и извлечение до 64 кадров на видео; кадры менее 256 px отбрасывались. Для фотодатасета из каждого видео извлекалось 5 равномерно распределённых кадров с дополнительной нормализацией ориентации лица по ключевым точкам (MediaPipe landmarks).

4. Предлагаемый метод

4.1 Фреймовые классификаторы

DINOv2-фото. Признаковый экстрактор — DINOv2 ViT-S/14 [8] с 22.1 млн параметров, предобученный самонаблюдаемым методом на датасете LVD-142M. Входное изображение масштабируется до 518×518 пикселей. Поверх бэкбона добавляется трёхслойная MLP-голова: 384→256→128→2 с ReLU и Dropout(0.3/0.2). Стратегия обучения двухфазная: первые 4 эпохи обучается только голова (бэкбон заморожен), с эпохи 5 размораживаются последние 3 блока трансформера. Оптимизатор — AdamW, функция потерь — CrossEntropyLoss; скорость обучения головы в 10 раз выше бэкбона (1e-4 и 1e-5 соответственно).

MobileNetV4-фото. Бэкбон mobilenetv4_conv_medium.e500_r256_in1k (~9.7 млн параметров, вход 224×224). Аналогичная двухфазная стратегия: заморозка на эпохах 0–2, разморозка с эпохи 3, LR 1e-4 / 1e-3.

4.2 Видеоклассификаторы

Для видеоуровня применяется единая схема: (1) равномерная выборка T = 16 кадров из видео (при нехватке — дублирование последнего кадра); (2) применение бэкбона к каждому кадру независимо, получение матрицы признаков [B, T, d]; (3) temporal mean pooling — усреднение по временно́й оси в вектор [B, d]; (4) MLP-голова: для DINOv2 — 384→256→128→2, для MobileNetV4 — 960→512→256→2.

Для исключения утечки данных разбивка train/val выполняется на уровне видео: все кадры одного видео попадают только в одну подвыборку. В качестве аугментаций применялись горизонтальное отражение (p = 0.5), поворот (±10°), ColorJitter (±0.2), Gaussian Blur и Random Sharpness. Одни и те же параметры применялись ко всем кадрам видео, чтобы не создавать искусственных межкадровых различий.

4.3 Детали обучения

Все эксперименты проводились на GPU NVIDIA GeForce RTX 3090 (24 ГБ VRAM) с использованием PyTorch Lightning, смешанной точностью AMP 16-bit, накоплением градиентов (4 шага) и ранней остановкой (patience = 5 по val_roc_auc). Batch size: 4 для DINOv2 (вход 518^2) и 16 для MobileNetV4 (вход 224^2). Максимальное число эпох — 50 для всех моделей.

5. Результаты экспериментов

5.1 Фреймовые классификаторы

На тестовом наборе из 9733 фотографий (Real: 3 084, Fake: 6 649) результаты представлены в таблице 2.

Таблица 2. Результаты фреймовых классификаторов (порог = 0.5)

Метрика	DINOv2-фото	MobileNetV4-фото
Полнота Real	99.51%	98.10%
Полнота Fake	99.73%	95.30%
val ROC-AUC (лучшая эпоха)	0.999	0.988

DINOv2 демонстрирует более высокое преимущество, особенно на сложных источниках: F++_FaceSwap, Dataset_Dasha1_sdxl, F++_FaceShifter. Артефакты диффузионной генерации и современных алгоритмов замены лиц фиксируются механизмом глобального внимания трансформера, но плохо улавливаются локальными свёрточными фильтрами.

5.2 Видеоклассификаторы

На тестовой выборке из 5417 видео получены результаты (таблица 3). Детализация по ключевым источникам для лучшей модели приведена в таблице 4.

Таблица 3. Результаты видео классификаторов на тестовой выборке

Метрика	DINOv2-видео	MobileNetV4-видео
ROC-AUC	0.9988	0.9932
Общая точность	98.30%	96.64%
Точность (Fake)	97.74%	96.73%
Точность (Real)	97.47%	96.49%

Таблица 4. Точность DINOv2 по ключевым источникам (тестовая выборка)

Источник	Класс	DINOv2-фото точность	DINOv2-видео точность
Real_0	Real	98.81%	91.55%
DFDC_real	Real	100.00%	93.33%
Real_1	Real	98.48%	98.15%
F++	Real	97.62%	98.00%
Real_Dasha	Real	97.06%	99.26%
DFD_original sequences	Real	100.00%	91.89%
Real_Dasha1	Real	100.00%	100.00%
Real_dive_photo	Real	100.00%	-
Real_new_videos	Real	100.00%	99.60%
Celeb_fakes	Fake	100.00%	98.95%
F++_FaceShifter	Fake	98.07%	94.00%
F++_Deepfakes	Fake	99.61%	100.00%
F++_FaceSwap	Fake	98.04%	100.00%
Dataset_Real_1_inswapper	Fake	98.05%	98.15%

DFD_manipulated_sequences	Fake	100.00%	97.97%
Dataset_Dasha1_sdxl	Fake	95.83%	100.00%
Dataset_Dasha_inswapper	Fake	97.22%	100.00%
DFDC_fake	Fake	100.00%	95.78%

Наибольшая доля ошибок у видеомодели по видео — на источниках Real_0 (91.55%) и Real/DFD_original (91.89%): видео Real_0 использовались как источник для SDXL-синтеза, поэтому их визуальные условия съёмки совпадают с синтетическими парными образцами; видео DFD_original подверглись многократному перекодированию, порождающему блочные JPEG-артефакты, схожие с признаками синтеза.

5.3 Сравнение архитектур

Трансформерный бэкбон (DINOv2) превосходит сверточный (MobileNetV4) во всех четырех постановках задачи. Наиболее выраженное преимущество — на фотоуровне: разрыв достигает 25 п.п. на отдельных источниках. На видеоуровне разрыв сокращается до 1.66 п.п.: временно́е усреднение по 16 кадрам компенсирует часть слабостей сверточного бэкбона.

6. Обсуждение

Обобщаемость. Признаки DINOv2, обученные самонаблюдаемым методом на разнообразных изображениях, хорошо переносятся на различные методы синтеза — в том числе на диффузионные модели (SDXL), отсутствовавшие в предобучении. Синтезированные SDXL лица распознаются с точностью 100% на видеоуровне: характерные высокочастотные артефакты диффузионного процесса хорошо фиксируются трансформерным вниманием. Это свойство особенно ценно в условиях постоянного появления новых методов генерации. В целом mean pooling по кадрам

игнорирует темпоральную динамику лица (мимику, движения головы), которая сама по себе является признаком синтеза — перспективным направлением остаётся применение Transformer-энкодера над последовательностью признаков.

7. Заключение

В работе представлена двухуровневая система обнаружения дипфейков на основе бэкбонов DINOv2 и MobileNetV4, обученная на гетерогенном датасете из более 43000 видеозаписей. Модель DINOv2 достигает ROC-AUC = 0.9988 на видеоуровне и ROC-AUC = 0.999 на фотоуровне, превосходя MobileNetV4 во всех постановках задачи. Сравнение подтверждает преимущество трансформерных архитектур с самонаблюдаемым предобучением для обнаружения дипфейков. Дальнейшие исследования направлены на разработку темпоральных моделей, учитывающих межкадровую динамику лица, и улучшение обобщаемости на сильно сжатых видео.

Список литературы

1. Dolhansky B., Bitton J., Pflaum B. [et al.] The DeepFake Detection Challenge (DFDC) dataset // arXiv:2006.07397. — 2020.
2. Haliassos A., Vougioukas K., Petridis S., Pantic M. Lips don't lie: A generalisable and robust approach to face forgery detection // Proc. CVPR. — 2021. — P. 5039–5049.
3. Li Y., Chang M.-C., Lyu S. In ictu oculi: Exposing AI created fake videos by detecting eye blinking // Proc. IEEE WIFS. — 2018. — P. 1–7.
4. Li Y., Yang X., Sun P. [et al.] Celeb-DF: A large-scale challenging dataset for deepfake forensics // Proc. CVPR. — 2020. — P. 3207–3216.

5. Mirsky Y., Lee W. The creation and detection of deepfakes: A survey // ACM Computing Surveys. — 2021. — Vol. 54, № 1. — P. 1–41.
6. Oquab M., Darcet T., Moutakanni T. [et al.] DINOv2: Learning robust visual features without supervision // Transactions on Machine Learning Research. — 2024.
7. Rossler A., Cozzolino D., Verdoliva L. [et al.] FaceForensics: A large-scale video dataset for forgery detection in human faces // arXiv:1803.09179. — 2018.
8. Rossler A., Cozzolino D., Verdoliva L. [et al.] FaceForensics++: Learning to detect manipulated facial images // Proc. ICCV. — 2019. — P. 1–11.
9. Tolosana R., Vera-Rodriguez R., Fierrez J. [et al.] Deepfakes and beyond: A survey of face manipulation and fake detection // Information Fusion. — 2020. — Vol. 64. — P. 131–148.
10. Verdoliva L. Media Forensics and DeepFakes: An Overview // IEEE Journal of Selected Topics in Signal Processing. — 2020. — Vol. 14, № 5. — P. 910–932.
11. Zhao H., Zhou W., Chen D. [et al.] Multi-attentional deepfake detection // Proc. CVPR. — 2021. — P. 2185–2194.
12. Zheng J., Bao W., Chen D. [et al.] Exploring temporal coherence for more general video face forgery detection // Proc. ICCV. — 2021. — P. 15044–15054.

References

1. Dolhansky B., Bitton J., Pflaum B. [et al.] The DeepFake Detection Challenge (DFDC) dataset // arXiv:2006.07397. 2020.
2. Haliassos A., Vougioukas K., Petridis S., Pantic M. Lips don't lie: A generalisable and robust approach to face forgery detection // Proc. CVPR. 2021. P. 5039–5049.

3. Li Y., Chang M.-C., Lyu S. In ictu oculi: Exposing AI created fake videos by detecting eye blinking // Proc. IEEE WIFS. 2018. P. 1–7.
4. Li Y., Yang X., Sun P. [et al.] Celeb-DF: A large-scale challenging dataset for deepfake forensics // Proc. CVPR. 2020. P. 3207–3216.
5. Mirsky Y., Lee W. The creation and detection of deepfakes: A survey // ACM Computing Surveys. 2021. Vol. 54, No. 1. P. 1–41.
6. Oquab M., Darcet T., Moutakanni T. [et al.] DINOv2: Learning robust visual features without supervision // Transactions on Machine Learning Research. 2024.
7. Rossler A., Cozzolino D., Verdoliva L. [et al.] FaceForensics: A large-scale video dataset for forgery detection in human faces // arXiv:1803.09179. 2018.
8. Rossler A., Cozzolino D., Verdoliva L. [et al.] FaceForensics++: Learning to detect manipulated facial images // Proc. ICCV. 2019. P. 1–11.
9. Tolosana R., Vera-Rodriguez R., Fierrez J. [et al.] Deepfakes and beyond: A survey of face manipulation and fake detection // Information Fusion. 2020. Vol. 64. P. 131–148.
10. Verdoliva L. Media Forensics and DeepFakes: An Overview // IEEE Journal of Selected Topics in Signal Processing. 2020. Vol. 14, No. 5. P. 910–932.
11. Zhao H., Zhou W., Chen D. [et al.] Multi-attentional deepfake detection // Proc. CVPR. 2021. P. 2185–2194.
12. Zheng J., Bao W., Chen D. [et al.] Exploring temporal coherence for more general video face forgery detection // Proc. ICCV. 2021. P. 15044–15054.