

ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ

Факултет „Компютърни системи и технологии“

Катедра „Компютърни системи“

Маг. инж. Георги Янев Котов

МЕТОДИ ЗА РАЗГРАНИЧАВАНЕ И РАЗПОЗНАВАНЕ НА ИЗОБРАЖЕНИЯ НА ЧОВЕШКИ ЛИЦА, ГЕНЕРИРАНИ С ИЗКУСТВЕН ИНТЕЛЕКТ, ОТ ИЗОБРАЖЕНИЯ НА РЕАЛНИ ЧОВЕШКИ ЛИЦА

А В Т О Р Е Ф Е Р А Т

**на дисертация за придобиване на образователна и научна степен
"ДОКТОР"**

Област: 5. Технически науки

Професионално направление: 5.3 Комуникационна и компютърна техника

**Научна специалност: „Автоматизирани системи за обработка на
информация и управление“**

Научни ръководители:

проф. д-р инж. Огнян Наков Наков

проф. д-р инж. Милена Кирилова Лазарова–Мицева

СОФИЯ, 2026 г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Компютърни системи“ към Факултет „Компютърни системи и технологии“ на ТУ–София на редовно заседание, проведено на 22.06.2025 г.

Публичната защита на дисертационния труд ще се състои на 29.09.2026 г. от 15,00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед № ОЖ-5.3-61 от 17.07.2026г. на Ректора на ТУ–София в състав:

1. проф.д-р Румен Иванов Трифонов – председател
2. доц.д-р Антония Тодорова Ташева – научен секретар
3. проф.д-р Атанас Велков Атанасов
4. проф.д-р Емил Иванов Йончев
5. доц.д-р Ивайло Симеонов Ченчев

Рецензенти:

1. проф.д-р Румен Иванов Трифонов
2. проф.д-р Атанас Велков Атанасов

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет „Компютърни системи и технологии“ на ТУ–София, блок №1, кабинет №1443а.

Дисертантът е редовен докторант към катедра „Компютърни системи“ на Факултет „Компютърни системи и технологии“. Изследванията по дисертационната разработка са направени от автора, като някои от тях са подкрепени от научноизследователски проекти.

Автор: маг. инж. Георги Котов

Заглавие: Методи за разграничаване и разпознаване на изображения на човешки лица, генерирани с изкуствен интелект, от изображения на реални човешки лица

Тираж: 30 броя

Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

В съвременното информационно общество лесно достъпният генеративен изкуствен интелект (ИИ) подкопава доверието във визуалните доказателства. Синтетичните изображения на човешки лица, произведени от генеративни състезателни мрежи (GAN), дифузионни модели, сред които Stable Diffusion, Imagen и DALL·E, или чрез техники за размяна на лице, вече представляват реален оперативен риск за цифрова идентичност, биометрична автентикация, медийна верификация и правоприлагане [1], [2], [3], [4].

Ключово форензично разграничение в настоящия труд е разграничаването между изображения с разменено лице и напълно ИИ-генерирани лица. При изображенията с разменено лице върху съществуващ кадър остават локални артефакти по границите на смесването, докато напълно синтетичните портрети могат да изглеждат глобално когерентни, но носят статистически аномалии на ниво текстура, симетрия и спектър [2]. Методите за откривателство, при които всички „фалшиви“ изображения се третират като един клас, губят форензична информация, необходима за модерация и разследване. Дисертационният труд адресира тези пропуски чрез интегрирана методология, обхващаща едновременно точност, адверсариална устойчивост, обясним ИИ и обобщаемост към невидяни генератори [5], [6], [7].

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на научните изследвания в дисертационния труд е да се предложат, изследват и оценят методи за разграничаване на ИИ-генерирани лица и лица с разменена идентичност в рамките на три-класова форензична таксономия с висока точност, устойчивост при L_∞ white-box атаки, последваща обяснимост и обединение в рамката RED-MS.

За постигането на поставената цел са изпълнени следните задачи:

- Да се извърши анализ на съществуващото състояние на генеративните модели за синтез на лица, методите за тяхното форензично разпознаване, адверсариалното машинно обучение и обяснимия ИИ, като се формулират изследователски пропуски и задачи на дисертацията.
- Да се разработи и оцени хибриден двукласов модел за откривателство, при който CNN семантичен признаков вектор (ResNet-18) се комбинира с форензичен white-pixel ratio дескриптор, върху набора WhichFaceIsReal.
- Да се разработи и оцени три-класова многопоточна архитектура с ViT, CNN и FFT потоци и сливане чрез внимание за разграничаване на реални изображения, изображения с разменено лице и напълно ИИ-генерирани изображения.
- Да се проектира и оцени рамката RED с FGSM/PGD адверсариално обучение, метрика ARI и Grad-CAM обяснимост.
- Да се синтезира методологията RED-MS, обединяваща многопоточното три-класово откривателство с адверсариалното укрепване и XAI в унифициран YAML-управляван конвейер, като се извърши сравнителен анализ и OOD оценка.
- Да се формулират научните приноси, ограниченията, етичните и правните аспекти, както и перспективите за развитие на изследването.

Научните методи, използвани в дисертационния труд, включват проучване и сравнителен анализ на литературни източници в областта; проектиране на невронни архитектури от вида ResNet-18, Xception, ResNet-50 и ViT-B/16 с PyTorch при стратифицирано разделяне на данните (70/15/15 и 80/20); L_∞ -ограничена адверсариална оценка при пълен достъп с FGSM и PGD при $\epsilon = 4/255$ и $8/255$; последваща обяснимост чрез Grad-CAM [90], [92].

Научна новост

Научната новост се изразява в последователното интегриране на четири методологични слоя, всеки от тях с документиран количествени резултати:

- Хибридният двукласов модел, при който ResNet-18 семантичен признаков вектор се комбинира с white-pixel ratio и класификатор тип случайни гори, постига 94.8% точност и $F1 = 0.942$, с документирано превъзходство над самостоятелното CNN (91.4% / 0.904) и самостоятелния форензичен подход (77.1% / 0.755).
- Три-класовата многопоточна архитектура, при която ViT-B/16, CNN и FFT потоци се сливат чрез механизъм за внимание, постига 93.7% обща точност, $F1$ по класове 0.97 за реалното, 0.90 за разменното лице и 0.94 за ИИ-генерираното изображение, $MCC \approx 0.90$ и $AUC (1-vs-all) > 0.98$.
- Рамката RED, при която FGSM/PGD адверсариалното обучение се интегрира с метриката ARI и Grad-CAM обяснимост, постига чиста точност Xception 97.53% и ResNet-50 96.64% при референтен чист протокол.
- Обединената архитектура RED-MS, при която многопоточността, адверсариалната защита и XAI се синтезират в единен YAML-управляван конвейер, постига при верифицирано PGD изпълнение чиста точност 98.14%, $macro-F1 = 0.9813$, $MCC = 0.9720$ и ARI при PGD-10 = 0.844.

Практическа приложимост

Предложените решения имат пряка приложна стойност в областта на медийна верификация, форензичен анализ и биометрична сигурност. Рамките RED и RED-MS предоставят методология за помощно откривателство, при което човек участва в цикъла на вземане на решения, без да се претендира за заместване на журналистическа проверка или правна квалификация. Реализираната отворена имплементация, YAML-конфигурируемият конвейер и документираните ограничения осигуряват практическа основа за възпроизвеждане и доразвиване от изследователски и приложни екипи, включително в контекста на националната инициатива за изследователска инфраструктура.

Апробация

Резултатите от дисертационния труд са разглеждани, обсъждани и публикувани в Международна научна конференция ICEST 2025, Охрид, Северна Македония (Scopus и WoS), Международна научна конференция COMSCI 2025, Созопол, България (Scopus), Международна научна конференция TechSys 2026, Пловдив, България (предстоящо публикуване в MDPI Engineering Proceedings, Scopus и WoS), Международна научна конференция COMSCI 2026, Созопол, България (приета за публикуване), Научен семинар на докторанти към катедра „Компютърни системи“, Факултет „Компютърни системи и технологии“, Технически университет – София.

Публикации

Основните постижения и резултати от дисертационния труд са публикувани в 4 научни статии. Научните статии са представени и публикувани в международни реферирани и индексирани издания и конференции (Scopus, WoS).

Структура и обем на дисертационния труд

Дисертационният труд е в обем от 127 страници като включва увод, четири глави за решаване на формулираните основни задачи, заключение, списък на основните приноси, списък на публикациите по дисертацията и използвана литература. Цитирани са общо 123 литературни источника. Трудът включва 28 фигури, 5 основни таблици и 4 алгоритма.

Номерата на фигурите, таблиците и използваната литература в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

УВОД

В днешно време лесно достъпният генеративен изкуствен интелект създава реален оперативен риск чрез синтетични човешки лица, произведени от GAN, дифузионни модели или face-swar инструменти. Централното форензично разграничение в настоящия труд е между изображения с разменено лице, при които върху съществуващ кадър остават локални артефакти по границите на смесването, и напълно ИИ-генерирани лица, при които глобалната когерентност е висока, но са налице аномалии на ниво текстура, симетрия и спектър. Изграждането на доверими модели за откривателство, устойчиви при адверсариални смущения и обясними чрез локализация на доказателства, е ключова предпоставка за верификация с участие на човек в цикъла.

Обект на изследването са методите за разграничаване на три форензично различни категории изображения на лица. Предмет на дисертационния труд са архитектурите и рамките за устойчиво и обяснимо откривателство на синтетични лица, при което три-класовата таксономия, адверсариалното обучение и Grad-CAM обяснимостта се разглеждат в унифициран конвейер. Работната хипотеза е, че комбинирането на семантични, локални и честотни признаци заедно с PGD адверсариалното обучение може да постигне висока точност в три-класовата постановка при запазена устойчивост, измерена чрез ARI.

ГЛАВА 1. АНАЛИЗ НА СЪЩЕСТВУВАЩОТО СЪСТОЯНИЕ. ДЕФИНИРАНЕ НА ПОСТАВЕНИТЕ ЗАДАЧИ. КОНЦЕПТУАЛНИ МОДЕЛИ

Главата преглежда съвременното състояние при откриване на синтетични и манипулирани лица, формализира терминологията и изследователските пропуски, които насочват Глави 2-4, и дефинира шестте изследователски задачи на дисертационния труд. Главата следва изследователската траектория на автора: хибриден двукласов модел за откриване, последван от три-класова многопоточна архитектура, адверсариално обучение и обяснимост (RED), и накрая концептуален синтез RED-MS.

1.1. Мотивация и обхват на прегледа

Синтетичните изображения на човешки лица вече не са академичен експеримент, а оперативен риск за цифрова идентичност, биометрична автентикация, медийна верификация и правоприлагане. Генеративните модели, от ранните GAN до съвременните латентни дифузионни модели, позволяват фотореалистични портрети с минимални технически умения, докато техниките за размяна на лице манипулират съществуващи изображения, запазвайки поза и изражение, но подменяйки самоличността. Двете категории не са форензично еквивалентни, което е централно за предложената три-класова форензична таксономия. Прегледът обхваща статични изображения на лица; откриването на ниво видео и биометричните сигнали (rPPG) са документиран като бъдещи разширения, но не са част от емпиричния обхват. Твърденията за адверсариална устойчивост в следващите глави са ограничени до L_∞ -ограничени FGSM и PGD атаки при пълен достъп (white-box).

1.2. Генеративни модели за синтез на лица

1.2.1. Генеративни състезателни мрежи (GAN)

Генеративните състезателни мрежи (GAN) [9] въвеждат състезателна парадигма генератор-дискриминатор. Ранните DCGAN [10] стабилизират обучението, а условните и

машабирани варианти, сред които BigGAN [11], CycleGAN [12] и StarGAN [13], разширяват контрола върху атрибутите. Прогресивното машабиране на ProGAN [14] и StyleGAN семейството [15], [16], [17] довеждат синтеза на лица до високо качество с контролирана вариация на поза, осветление и израз. С развитието на генераторите разликата в артефактите се стеснява, което мотивира по-богати, хибридни и многопоточни представяния на признаците.

1.2.2. Дифузионни модели и латентна дифузия

Дифузионните модели (DDPM [19] и техните усъвършенствания [20], [21]) моделират генерацията като постепенно премахване на шум и често постигат по-стабилно обучение и висок фотореализъм. Латентните дифузионни модели прилагат дифузията в компресирано латентно пространство, което позволява обуславяне върху текст, така че Stable Diffusion [24], Imagen [25], [26], DALL·E и DALL·E 2 [27], [28] демократизират синтеза от текст към изображение. GAN изходите носят по-явни високочестотни и периодични артефакти, докато дифузионните изходи са по-гладки, с отклонения на ниво разпределение, което е различие, релевантно за избора между CNN и ViT потоци.

1.2.3. Размяна на лица и класически процеси

Изображенията с разменено лице представляват отделен клас манипулация: не се генерира ново лице, а се трансплантира самоличност от изходно към целево изображение чрез инструменти като Face2Face [32], DeepFaceLab [33] и Faceswap [34]. Артефактите са предимно локални, а именно граници на смесване, несъответствие в цвета и времево трептене. Форензичната литература [35] настоява изображенията с разменено лице и ИИ-генерираните лица да се третират като различни модели на заплаха, което е методологична предпоставка за три-класовата класификация и за обединената рамка RED-MS.

1.2.4. VAE и спомагателни генеративни парадигми

Вариационните автокодери (VAE) и нормализиращите потоци моделират латентни разпределения с явни функции на правдоподобие и често служат като компонент в латентните дифузионни модели, а не като самостоятелен фотореалистичен генератор. За откриването релевантността е косвена, докато основният фокус на дисертацията остава върху моделите на заплаха GAN/дифузия/размяна, доминиращи в публичните еталонни тестове.

1.2.5. Еволюция на генеративната заплаха (хронология)

Очертава се хронология: 2014-2017 г., когато се появяват GAN и ранни размени на лица; 2018-2020 г., когато StyleGAN и CNN моделите стават „лесни за разпознаване“; 2020-2022 г., когато се появяват StyleGAN2/3 и многопоточни модели; и периодът от 2022 г. до настоящия момент, в който LDM/Stable Diffusion, Imagen и DALL·E пораждат нужда от устойчиви и обясними обединени рамки (RED, RED-MS). Тази еволюция мотивира непрекъснато преобучение и оценка между генератори, а не еднократен еталонен тест.



Фиг. 4. Историческа хронология.

1.3. Заплахата от синтетично лице и синтетични лица

1.3.1. Социален и оперативен контекст

Технологиите за синтетични лица намират приложение в измама със самоличност, фишинг, атаки срещу репутацията и заобикаляне на KYC проверки [36], [37], [38]. Заплахата е асиметрична, тъй като разходът за генериране намалява, докато разходът за надеждна верификация остава висок. Моделите за откриване трябва да работят и при JPEG компресиране, преоразмеряване и филтри на социалните мрежи, които представляват верига от изкривявания, документирана в WildRF [39] и Deepfake-Eval-2024 [40].

1.3.2. Три-класова форензична таксономия

За целите на дисертацията се формализират три форензично различни категории: реални изображения (автентично заснето лице без манипулация); изображения с разменено лице (типични артефакти, а именно граници на смесване и локални несъответствия, при представителни източници FaceForensics++ и DeepFaceLab); ИИ-генерирани лица (глобални аномалии в текстура и симетрия, спектрални сигнали от генератори като StyleGAN и Stable Diffusion, набор от данни AI-Face). Двете категории фалшификации не се обединяват в един „фалшив“ клас.

1.3.3. Ограничения на точността като метрика

Висока точност върху отделен тестов набор не имплицира оперативна надеждност, тъй като моделите могат да се пренастроят към специфики на набора, например модел на камерата, ниво на компресия или конкретна версия на генератора, използвани при събирането на данните. Демографските отклонения и пропуските в безпристрастността, документиращи в AI-Face, мотивират експерименти по обобщаемост, разгледани в етичния подраздел на заключението. Допълнително, точността е чувствителна към дисбаланса между класовете: при силно небалансирани набори от данни висока обща точност може да маскира системно ниска ефективност върху по-рядко срещания клас, поради което дисертацията последователно допълва точността с F1-мярка, MCC и AUC-ROC по класове, а не разчита на единичен обобщен показател.

Отделно, точността, измерена при еднократно случайно разделяне на данните, не отразява вариацията, която произтича от избора на начална стойност (seed), реда на партидите по време на обучение или конкретната стратификация на разделянето. По тази причина дисертацията фиксира seed = 42 за всички сравними експерименти и документира протокола на разделяне, за да направи докладваните стойности възпроизводими и сравними между конфигурациите, а не статистически случайни наблюдения.

1.4. Таксономия на методи за откриване на синтетични и манипулирани лица

1.4.1. Еднопоточни методи

CNN основни мрежи (ResNet, Xception, EfficientNet) доминират откриването на синтетични лица на ниво изображение благодарение на способността да улавят йерархични локални текстури. Ранни специализирани модели като MesoNet [46] и методи, акцентирани върху границите на смесване (Face X-ray), показват, че фокусът върху локални несъответствия подобрява обобщаемостта, но единичен поток без явни честотни или форензични сигнали може да пропусне манипулации или да подцени дифузионни изходи.

1.4.2. Хибридни подходи

Хибридните подходи комбинират научени семантични признаци с ръчно проектирани форензични дескриптори, например признаков вектор от ResNet-18 в комбинация с white-pixel ratio, подадени на класификатор Random Forest. Това подобрява двукласовото

откриване спрямо едноточните подходи, но сливането често остава частично непрозрачно (black-box), което мотивира по-структурирано многопоточно сливане чрез внимание.

1.4.3. Многопоточни и мултимодални архитектури

Многопоточните модели обединяват паралелно ViT-B/16 (глобална кохерентност и симетрия), CNN (Xception/ResNet, локални артефакти на смесване) и FFT (спектрални нередности). Сливането чрез внимание динамично претегля приноса на всеки поток според типа манипулация и реализира три-класовата форензична таксономия.

1.4.4. Форензични сигнали и анализ в честотната област

Литературата идентифицира повтарящи се форензични сигнали: граници на смесване, несъответствия в лицевите точки, спектрални аномалии в 2D FFT/DCT, бинарно съотношение на бели пиксели, взаимодействие с JPEG компресията и нередности в цветовото пространство [41], [42]. Тези сигнали мотивират FFT потока в RED-MS, но са нестабилни при адверсариални смущения, което е мотивация за адверсариално обучение.

1.4.5. Vision Transformers и ограничения

ViT прилага внимание тип transformer и теоретично улавя глобални несъответствия, но в RED постига само 69.08% точност при Kaggle набор, което е резултат, специфичен за протокола (10 епохи, ограничен бюджет), а не общ отказ на архитектурата. Обяснимостта на ViT изисква разгъване на вниманието (attention rollout) или AttnLRP, а не стандартен Grad-CAM.

1.4.6. Метрики за оценка на модели за откриване

Използва се стандартен набор от метрики: Accuracy, Precision/Recall/F1 по класове, Matthews Correlation Coefficient (MCC), AUC-ROC (един срещу всички), матрица на обърквания, както и метриката ARI (Adversarial Robustness Index) за адверсариална оценка и EF за обяснимост, дефинирани формално в раздел 1.6.

1.4.7. Обобщени обзорни перспективи

Прегледите по медийна форензика [45] класифицират методите по архитектура, тип манипулация и модалност. Биологичните и поведенческите сигнали са комплементарни, но извън обхвата на статичните изображения. Литературата за премахване на отклонения на ниво данни подчертава нуждата от демографска осведоменост и тестове извън разпределението (OOD).

1.5. Адверсариално обучение и устойчивост

1.5.1. Адверсариални примери и модел на заплахата

Адверсариалните примери са входове, получени чрез малка пертурбация δ , която променя изхода на класификатора. FGSM изчислява δ с една стъпка по знака на градиента: $x' = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{A}(\theta, x, y))$. PGD итеративно усъвършенства пертурбацията в L_∞ : $x_{n+1} = \Pi(x_n + \alpha \cdot \text{sign}(\nabla_x \mathcal{A}(\theta, x_n, y)))$. Madry и съавтори [92] формализират устойчивата оптимизация чрез PGD като стандарт за адверсариално обучение, при типични бюджети $\epsilon = 4/255$ и $8/255$.

1.5.2. Баланс между точност и устойчивост при адверсариално обучение

Адверсариалното обучение минимизира загубата в най-лошия случай по δ . Устойчивите модели често понижават незначително чистата точност, но повишават точността под атака. Адверсариалните оценки в дисертацията са ограничени до L_∞ FGSM/PGD при пълен достъп (white-box); атаките тип черна кутия, трансферните и адаптивните атаки се цитират като открити проблеми.

1.6. Метрики за оценка – формални дефиниции

Раздел 1.6 формализира класификационните метрики, а именно $Accuracy = (TP+TN)/(TP+TN+FP+FN)$; $Precision = TP/(TP+FP)$; $Recall = TP/(TP+FN)$; $F1 = 2 \cdot Precision \cdot Recall / (Precision + Recall)$; $MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{((TP+FP)(TP+FN)(TN+FP)(TN+FN))}$; AUC-ROC (един срещу всички), както и метриката $ARI = Acc_adv / Acc_clean$ за адверсариална устойчивост и EF (пространствена корелация на Grad-CAM с маска на манипулация) за обяснимост. Тези дефиниции се прилагат последователно в Глави 2-4.

1.7. Обясним ИИ (XAI) при форензично откриване

1.7.1. Мотивация за обяснимост при откриването на синтетични лица

Общественият натиск и политиките за модерация изискват не само класификационна оценка, но и обосновка къде моделът „вижда“ доказателства за фалшификация. Прегледът систематизира последващите XAI методи по аксиоматични и пертурбационни критерии.

1.7.2. Grad-CAM и обяснимост при CNN

Grad-CAM комбинира градиентите на целевия клас с активациите на последния конволюционен слой, за да получи топлинна карта на дискриминативните области. В RED Grad-CAM се прилага върху Xception и ResNet-50 за клас Fake и дава качествено съгласувани топлинни карти, но не е еталонна сегментация и е чувствителен към избора на слой.

1.7.3. Алтернативни XAI методи

Разглеждат се Layer-wise Relevance Propagation (LRP), Integrated Gradients, LIME, SHAP, RISE, както и разгъване на вниманието (attention rollout) и AttnLRP, които са специфични за ViT и необходими, когато Grad-CAM не се поддържа.

1.7.4. Обяснимост на предложения модел

EF количествено измерва пространственото припокриване между картата на значимост и еталонна маска на манипулация (напр. от FaceForensics++) чрез корелация на Pearson или Jaccard-подобно припокриване. В дисертацията EF се прилага качествено в RED и количествено се планира в RED-MS.

1.8. Данни, протоколи и еталонни тестове

Разделът обобщава протоколите за разделяне на данни и основните еталонни набори, използвани в Глави от втора до четвърта: разделяне по самоличност/видео/генератор за избягване на изтичане на данни (раздел 1.8.1) и основните набори от данни (раздел 1.8.2), а именно WhichFaceIsReal, FaceForensics++, AI-Face, AI-vs-Deepfake-vs-Real, Kaggle Real/Fake Faces, WildRF и Deepfake-Eval-2024, всеки с конкретна роля в експерименталните глави.

1.8.3. WhichFaceIsReal

Около 10 000 изображения, реални срещу StyleGAN-генерирани портрети, се използват за хибридни експеримент при разделяне 80/20. Ограничението е, че фокусът е само върху StyleGAN семейството.

1.8.4. AI-vs-Deepfake-vs-Real

Три-класов балансиран набор от 9 999 изображения; стратифицирано разделяне 70/15/15, което е стандарт за три-класовите експерименти и за RED-MS методологията.

1.8.5. FaceForensics++

Видеокадри с множество манипулации и пикселни маски на ниво пиксел, които служат като референтен набор за бъдещата количествена EF оценка в RED-MS.

1.8.6. AI-Face

Около един милион демографично анотирани ИИ/реални лица, използван за разширяване на разпределението при обучение и за бъдещ одит на безпристрастност.

1.8.7. Deepfake-Eval-2024

Синтетични лица, разпространени през 2024 г., с мултимодална анотация, които се използват само за тест извън разпределение (OOD).

1.8.8. WildRF

Набор от данни за синтетични лица в реална среда (in-the-wild), използван за zero-shot OOD тест без участие в обучението.

1.8.9. Kaggle real/fake faces (RED)

Бинарен набор със стратифицирано разделяне 70/15/15, който служи като опорна точка за протокола RED и е концептуално комплементарен на три-класовото разделяне.

1.9. Дефиниране на поставените задачи

В резултат на анализа се формират шест задачи пред дисертационния труд:

- Анализ на съществуващото състояние, формулиране на изследователските пропуски и дефиниране на задачите (настоящата глава).
- Разработка и оценка на хибриден двукласов модел за откриване (ResNet-18 и white-pixel ratio) върху WhichFaceIsReal (Глава 2).
- Разработка и оценка на три-класова многопоточна архитектура ViT, CNN и FFT със сливане чрез внимание (Глава 2).
- Проектиране и оценка на рамката RED с FGSM/PGD адверсариално обучение, метрика ARI и Grad-CAM обяснимост (Глава 3).
- Синтез на методологията RED-MS за обединение на многопоточното откриване с адверсариалното укрепване и XAI, сравнителен анализ и разширени експерименти (Глава 4).
- Формулиране на научните, научно-приложните и приложните приноси, обобщаване на изпълнението на задачите, ограниченията, етичните и правните аспекти и перспективите за развитие.

1.10. Концептуални модели и изходни точки

1.10.1. Модел на заплахата

Моделът на заплахата включва разпределение на синтетични и разменени лица от разнородни генератори, противник с пълен достъп (white-box) с ϵ -ограничени пертурбации и оператор (модератор/експерт), нуждаещ се от локализирани доказателства.

1.10.2. Модел на детектора като композиция от потоци

Обобщената функция $f(x) = \text{Classifier}(\text{Fusion}(\phi_{\text{ViT}}(x), \phi_{\text{CNN}}(x), \phi_{\text{FFT}}(x)))$ формализира многопоточната архитектура; RED добавя адверсариална загуба върху f , а RED-MS прилага същото върху три-класовата постановка.

1.10.3. Критерии за оценка

Критериите обхващат класификация (Accuracy, F1, MCC, AUC), устойчивост (ARI, Acc_adv), обяснимост (Grad-CAM качествено, EF) и обобщаемост (разлика извън разпределение, OOD).

1.11. Формализация на изследователската хипотеза

Формализирани са шест хипотези H1-H6: H1 предполага, че хибридизацията превъзхожда самостоятелните CNN и форензичен поток върху WhichFaceIsReal; H2 предполага, че три-класовият многопоточен класификатор постига $F1 > 0.85$ за всеки клас; H3 предполага, че адверсариалното обучение води до компромис между чиста и адверсариална точност, измерим чрез ARI; H4 предполага, че Grad-CAM активациите корелират визуално с форензичните области при CNN; H5 предполага, че точността извън разпределението е по-ниска от тази в разпределението за невидени генератори; H6 предполага, че RED-MS не деградира три-класовата чиста точност спрямо самостоятелния многопоточен модел при идентичен протокол.

1.12. Изводи към първа глава

Прегледът на съвременното състояние очертава контекста и обосновава задачите на дисертационния труд:

- Генеративните модели са достигнали фотореалистично качество, което прави визуалното откриване практически невъзможно без автоматизирани форензични методи.
- Методите за откриване се систематизират в допълващи се групи, а именно пространствени (CNN), глобално-контекстни (ViT), честотни (FFT) и пиксел-форензични; нито една отделна модалност не е достатъчна срещу пълния спектър от генератори.
- Адверсариалната уязвимост и липсата на обяснимост остават двата най-съществени пропуски за практическо внедряване.
- Обобщаемостта между невидени генератори (OOD) и количествената оценка на обяснимостта (EF) са недостатъчно изследвани и мотивират методологията в следващите глави.
- Дефинираните метрики (Acc, F1, MCC, AUC, ARI, EF) и хипотезите H1-H6 задават измеримите критерии за оценка на приносите на труда.

ГЛАВА 2. ПРЕДЛОЖЕНИ АРХИТЕКТУРИ ЗА ОТКРИВАНЕ НА СИНТЕТИЧНИ ЛИЦА: ХИБРИДЕН И МНОГОПОТОЧЕН ПОДХОД

В тази глава са представени две авторски архитектури за откриване на синтетични лица: хибриден двукласов модел, комбиниращ дълбоки CNN признаци с форензични пикселни дескриптори, и многопоточна три-класова архитектура със сливане чрез внимание. Двете разработки формират емпиричната основа за последващото разширяване към рамките RED и RED-MS.

2.1. Форензични следи на генеративните модели

Генеративните модели оставят характерни форензични следи: GAN изображенията проявяват периодични „шахматни“ артефакти и спектрални пикове от транспонирани конволюции; дифузионните изображения носят по-фини, разпределени отклонения; изображенията с разменено лице проявяват локални артефакти по границата на смесване. Тази онтологична разлика обосновава защо размяната и пълният синтез се третират като отделни форензични класове и защо откриването изисква едновременно глобален, локален и честотен анализ [41], [42], [103], [123].

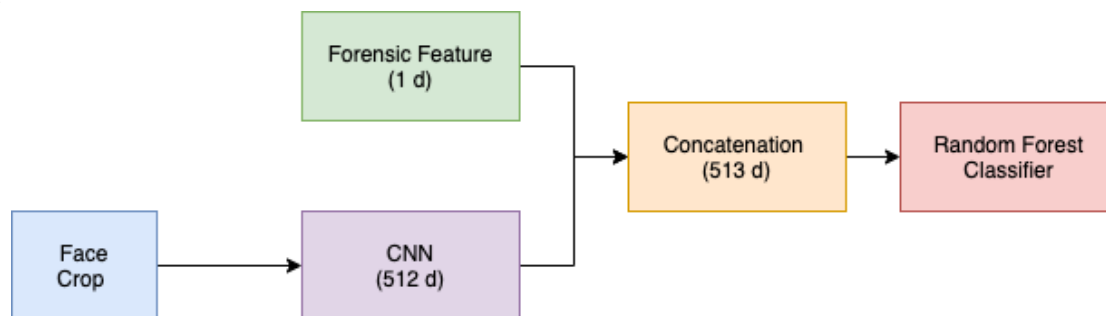
2.2. Хибриден подход: CNN и форензични признаци за двукласово откриване

2.2.1. Мотивация и постановка на задачата

CNN класификаторите улавят семантични структури, но могат да пропуснат тънки артефакти, докато форензичните евристики са интерпретируеми, но самостоятелно дават ограничена точност. Задачата проверява дали сливането им чрез конкатенация подобрява двукласовото откриване (реално срещу ИИ-генерирано) спрямо единичен поток [51].

2.2.2. Архитектура на хибридният процес

Методът интегрира 512-мерен семантичен признаков вектор от предварително обучен ResNet-18 [52] и форензичен скалар white-pixel ratio, изчислен след бинаризация на изрязаното лице. Двата вектора се конкатенират в 513-мерен вход, върху който се обучава RandomForestClassifier със 100 дървета при стратифицирано разделение 80/20 [53], [54], [55].



Фиг. 14. Хибриден подход за откриване на ИИ-генерирани лица [53]

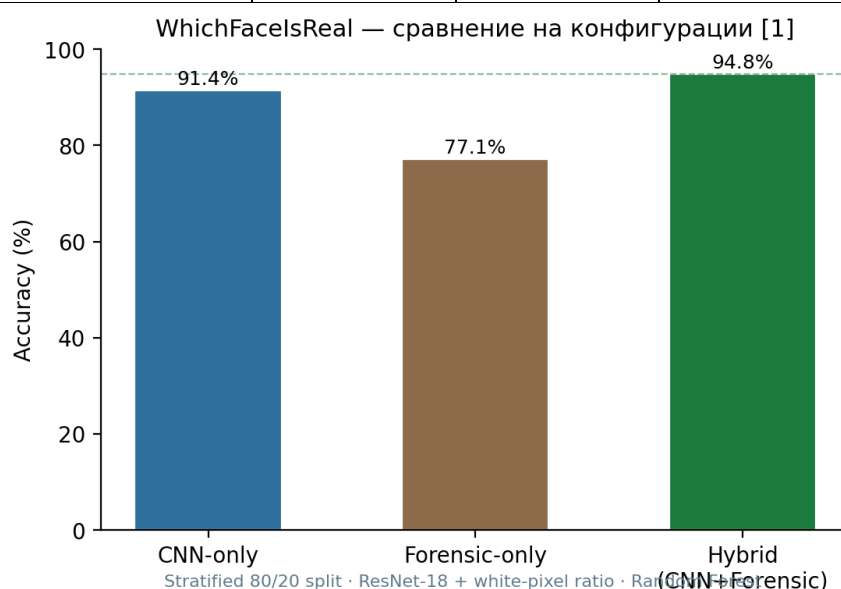
2.2.3. Експериментална методология

Оценката използва WhichFaceIsReal, който съдържа 10 000 изображения и 12 000 предварително етикетирани лица, покриващи StyleGAN1/2/3 и StyleFaceV [56]. Сравняват се три конфигурации, а именно CNN-only, Forensic-only и Hybrid [57], [58], чрез Accuracy, Precision, Recall и F1-score върху тестовото подмножество (20%), с фиксирана начална стойност (seed) 42 [59].

2.2.4. Експериментални резултати

Таблица 2.1. Експериментални резултати върху обучителния/тестов набор (WhichFaceIsReal, стратифицирано разделяне 80/20)

Модел	Accuracy	Precision	Recall	F1 score
CNN-only	91.4%	90.2%	90.7%	90.4%
Forensic-only	77.1%	74.9%	76.2%	75.5%
Hybrid (CNN и Forensic)	94.8%	94.1%	94.3%	94.2%



Фиг. 15. Точност по конфигурация върху WhichFaceIsReal (създадена от автора въз основа на Таблица 2.1).

Хибридният модел постига 94.8% точност и F1 = 0.942, надминавайки CNN-only (+3.4 п.п.) и Forensic-only (+17.7 п.п.), което потвърждава комплементарността на семантичните и форензичните сигнали.

2.2.5. Ограничения

Изчислителната цена на двойното извличане на признаци увеличава латентността; внедряване в реално време не се претендира. Моделът е оценен само върху StyleGAN семейството, а обобщаемостта между демографски групи остава отворен въпрос.

2.3. Свързани подходи и позициониране

Предложените разработки се позиционират спрямо три установени посоки: единичните CNN модели за откриване, чисто честотните (форензични) модели и хибридните/многопоточните комбинации. Единичните CNN модели разчитат на имплицитно научени признаци, докато честотните модели улавят процесно-инвариантни спектрални подписи, но са слаби срещу локални артефакти на размяна. Хибридните и многопоточните подходи целят да съчетаят комплементарните силни страни на тези посоки; първата разработка прави това в минималистична двупризнакова форма, а втората разширява принципа до три специализирани потока с обучавемо сливане.

За разлика от подходи, които третират откриването на синтетични лица като еднороден бинарен проблем, настоящата работа изрично разграничава форензичните категории по техния произход – размяна на лице срещу пълен синтез – и проектира потоците на признаци съобразно тази таксономия. Това позициониране е съществено различно от литературата, която обикновено оптимизира единична метрика върху общ „фалшив“ клас, без да разглежда как различните типове манипулация влияят на интерпретацията на грешките. Освен това, докато повечето сравними изследвания докладват резултати върху единичен еталонен набор, настоящите разработки явно документират границите на обобщаемост между набори от данни и генератори, което е предпоставка за по-късното въвеждане на систематичен out-of-distribution протокол в Глава 4.

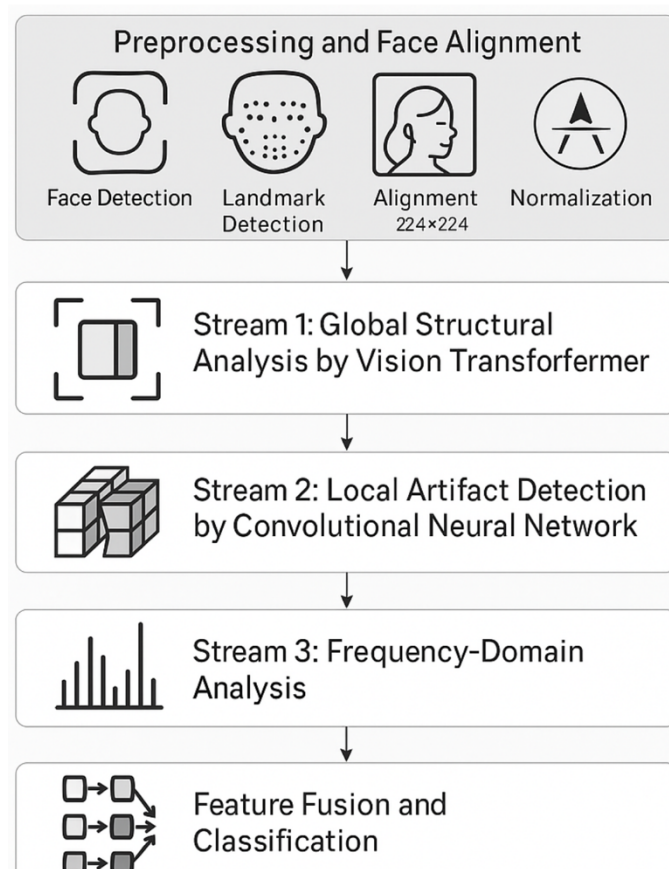
2.4. Многопоточни признакови вектори за три-класово откриване

2.4.1. Мотивация и постановка на задачата

Хибридният двукласов модел не разграничава изображенията с разменено лице от напълно синтетичните. Задачата изисква три-класов класификатор (реално / с разменено лице / ИИ-генерирано), при който всеки поток е настроен към различен профил на артефакти, а модулът за сливане динамично претегля потоците [61].

2.4.2. Архитектура на многопоточния процес

Методът интегрира три потока, а именно ViT-B/16 за глобална кохерентност, CNN (Xception/ResNet-50) за локални артефакти и 2D-FFT за спектрални аномалии, след стандартизирана предобработка (откриване на лице, подравняване, изрязване до 224×224 px) [62]. Признаковите вектори се сливат чрез лек transformer енкодер с механизъм за внимание вместо проста конкатенация, за да позволи динамично претегляне в зависимост от входа [63], [64], [65], [66].



Фиг. 16. Многопоточна архитектура за три-класов вектор на признаци.

2.4.3. Експериментална методология

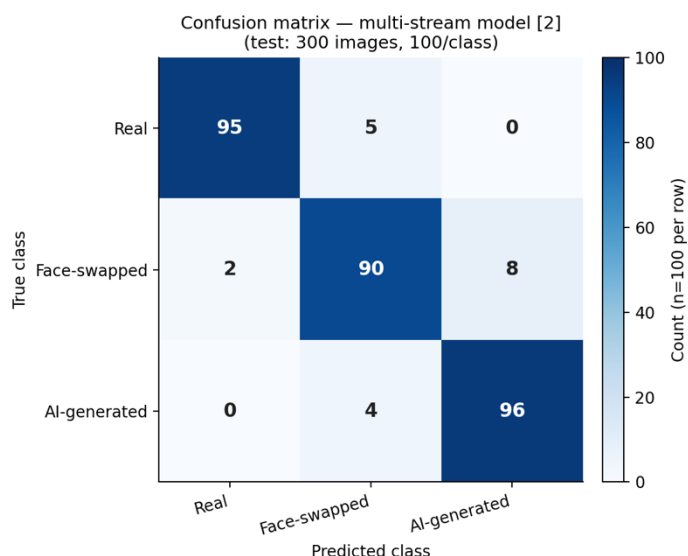
Моделът е обучен и оценен върху AI-Face, AI-vs-Deepfake-vs-Real, FaceForensics++ (обучение/валидация), и Deepfake-Eval-2024/WildRF (само тест) [67], [68], [69]. Разделянето е стратифицирано 70/15/15 с фиксиран seed 42. Оценъчните метрики включват Accuracy, Precision/Recall/F1 по класове, macro F1, one-vs-all AUC-ROC и MCC.

2.4.4. Експериментални резултати

Таблица 2.2. Класификационни резултати върху тестовия набор (300 изображения, 100/клас)

Клас	F1-Score
Истински	0.97
Face-swapped (Разменени лица)	0.90
ИИ-генерирано	0.94

Общата точност е 93.7% (281/300), MCC $\approx$ 0.90, one-vs-all AUC > 0.98 [70]. Матрицата на обърквания показва, че остатъчното объркване е концентрирано между разменено лице и ИИ-генерирано (8 и 4 гранични грешки), което потвърждава, че категориите са форензично различни, но не тривиално разделими.



Фиг. 18. Матрица на обърквания на многопоточния модел.

2.4.5. Обосновка на сливането чрез внимание

Сливането чрез внимание $h = \alpha_{v.e_ViT} + \alpha_{c.e_CNN} + \alpha_{f.e_FFT}$ (с $\sum \alpha_i = 1$) реализира сливане, обусловено от образа, за разлика от фиксираната конкатенация. То намалява риска от пренастройване и позволява модела да „пренасочи вниманието“ към доминиращия поток за конкретен тип манипулация.

2.4.6. Валидност на сравненията и граници на обобщение

Директното съпоставяне на 94.8% (2-class) и 93.7% (3-class) като подреждане по превъзходство е методологически невалидно, тъй като различни са наборите от данни, трудността на задачата и метриките. Двете стойности се третират като точки от обща изследователска траектория [71].

2.4.7. Ограничения

Трите паралелни основни мрежи изискват значителен обем GPU памет; около 8% от изображенията с разменено лице и 4% от ИИ-генерираните се объркват взаимно, а демографският одит за безпристрастност остава задача за бъдеща работа [72].

2.5. Дискусия

Двете разработки се разглеждат като последователни точки върху обща изследователска траектория, а не като конкуриращи се решения. Сравнението между 94.8% и 93.7% точност не е транзитивно поради различните набори, трудност на задачата и метрики [73]. Наблюдаваните гранични грешки между размяна на лице и пълнен синтез носят самостоятелна научна стойност и мотивират рамките RED и RED-MS [74].

2.6. Изводи към втора глава

- Хибридният двукласов модел (CNN + форензичен поток) постига 94.8% точност върху WhichFaceIsReal, над самостоятелните CNN (91.4%) и форензичен (77.1%) подход, потвърждавайки комплементарността на сигналите (хипотеза H1).
- Многопоточната три-класова архитектура постига 93.7% обща точност при F1 по класове 0.97/0.90/0.94, потвърждавайки, че разменното лице и пълният синтез са форензично различими класове (H2).
- Двете стойности не са транзитивно съпоставими, тъй като произлизат от различни набори, трудност на задачата и метрики.
- Структурираните грешки между двата фалшиви класа (~8%/~4%) мотивират добавянето на гранични примери и устойчивост към компресия.

- Основните ограничения, а именно изчислителната цена, чувствителността към ниска резолюция/компресия и невалидираният демографски одит, очертават обхвата на валидност и мотивират рамките RED и RED-MS.

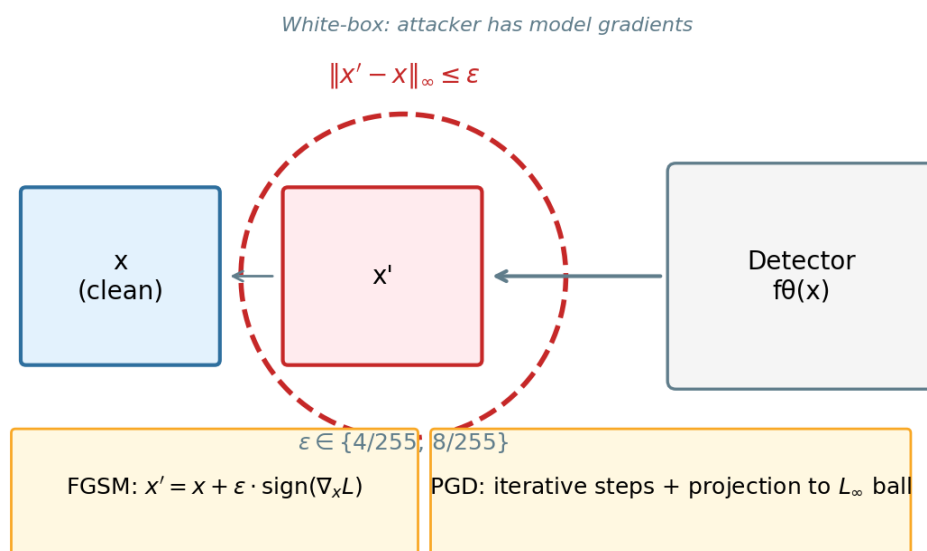
ГЛАВА 3. СИНТЕЗ НА ПРАВИЛА, МЕТОДИ И АЛГОРИТМИ ЗА АДВЕРСАРИАЛНА УСТОЙЧИВОСТ И ОБЯСНИМОСТ ПРИ ОТКРИВАНЕ НА ИИ-ГЕНЕРИРАНИ ЛИЦА

Главата представя авторската рамка RED (Robust and Explainable Detection), която обединява откриването на синтетични лица с адверсариално обучение (FGSM/PGD) и последваща обяснимост чрез Grad-CAM, в двукласова постановка преди обединението с многопоточната архитектура в RED-MS.

3.1. Мотивация и постановка на задачата

Чистата (clean) точност сама по себе си не гарантира, че моделът разчита на устойчиви форензични сигнали. RED изследва две взаимосвързани постановки: (а) влиянието на FGSM/PGD адверсариалното обучение върху устойчивостта при L_∞ пертурбации, количествено изразено чрез ARI, и (б) дали топлинните карти на Grad-CAM остават семантично съгласувани и стабилни при адверсариален вход [75], [76]. Твърденията са ограничени до L_∞ FGSM/PGD при пълен достъп (white-box), $\epsilon = 4/255$ и $8/255$.

L_∞ threat model: FGSM и PGD (white-box)



Фиг. 11. Постановка на адверсариална атака при пълен достъп (white-box) в обхвата на RED.

3.2. Архитектурен преглед на RED рамката

RED е модулна експериментална рамка с три етапа: извличане на признаци (CNN, а именно Xception и ResNet-50; ViT-B/16 за глобална кохерентност) чрез timm с ImageNet предобучение, адверсариално обучение чрез FGSM/PGD аугментация на мини-партидите, и последваща обяснимост чрез Grad-CAM върху последния конволюционен слой на CNN мрежите за клас Fake. Grad-CAM не е директно приложим към ViT-B/16, който изисква разгъване на вниманието.

3.2.1. Обосновка на избора на основни мрежи

Xception улавя локални текстурни и гранични артефакти; ResNet-50 предоставя широко използвана референтна точка; ViT-B/16 внася глобално внимание и служи за контраст и бъдеща обяснимост, базирана на вниманието. И трите използват трансферно

обучение от ImageNet [78], а адверсариалният защитен слой не въвежда нови параметри, а само модифицира разпределението на обучителните партии.

3.3. Адверсариално обучение и метрики за устойчивост

Адверсариалното машинно обучение моделира най-лошия случай на малка входна промяна. FGSM генерира една градиентна стъпка; PGD итерира многократно с проекция в ϵ [79]. И двата метода са с пълен достъп (white-box).

3.3.1. Таксономия на моделите на заплаха

Атаките се делят по достъп (white-box, black-box, трансферни [80], адаптивни [81]) и по цел (насочени/ненасочени). Геометрията на пертурбациите се ограничава чрез L_p -норма; RED ограничава всички твърдения до L_∞ -ограничени FGSM и PGD с пълен достъп при $\epsilon = 4/255$ и $8/255$, на скала $[0,1]$.

3.3.2. Постановка на оптимизацията

Адверсариалното обучение се формулира като вложена оптимизация: $\min_{\theta} E_{(x,y)}[\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(\theta, x+\delta, y)]$, апроксимирана с FGSM ($x' = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L})$) или PGD ($x_{(n+1)} = \Pi_{\epsilon}(x_n + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}))$), с комбинирана загуба $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_{\text{cons}}$.

3.3.3. Дълбочина на анализа: FGSM спрямо PGD

FGSM е евтина еднократна стъпка, подходяща за аугментация при обучение. PGD обобщава FGSM до итеративна процедура с N стъпки и случаен рестарт, който засилва атаката при оценка, така че тя представлява по-надеждна горна граница на white-box уязвимостта и диагностика срещу маскиране на градиента [83].

3.3.4. Протокол за обучение и оценка

Бюджетите са $\epsilon = 4/255$ и $8/255$; режимите $\text{adv_train} \in \{\text{none}, \text{fgsm}, \text{pgd}, \text{mix}\}$ управляват аугментацията. При оценка модулът извежда изход при условия clean, FGSM и PGD с еднакви ϵ за сравнимост между модели [84].

3.3.5. Алгоритмична формализация и сложност

Каноничните уравнения се обобщават в три алгоритъма: Алгоритъм 3.1 за адверсариално обучение (FGSM/PGD аугментация и AdamW актуализация), Алгоритъм 3.2 за оценка и изчисляване на ARI, и Алгоритъм 3.3 за Grad-CAM при извеждане на изход [85]. PGD-обучението с K стъпки изисква $O(K \cdot C)$ на партида, а многопоточният вариант RED-MS мащабира цената линейно с броя потоци S , т.е. $O(S \cdot C)$.

3.3.6. Компромис точност-устойчивост (качествен анализ)

Адверсариалното обучение принуждава модела да изостави неустойчиви, но силно предсказателни признаци в полза на по-стабилни сигнали, което подобрява поведението под атака, но може да понижи чистата точност. ARI винаги се интерпретира съвместно с $\text{Acc}_{\text{clean}}$, а не изолирано.

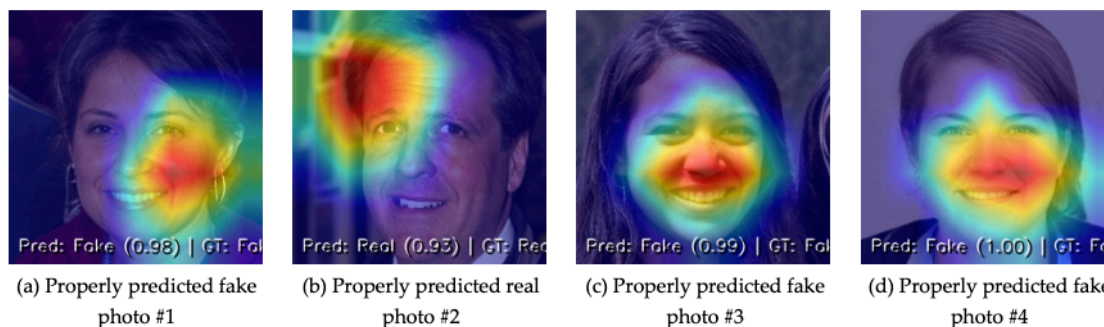
3.3.7. Adversarial Robustness Index (ARI)

$\text{ARI} = \text{Acc}_{\text{adv}} / \text{Acc}_{\text{clean}}$ количествено обобщава относителния спад при атака, отчитан поотделно за FGSM и PGD. Таблиците с най-добри валидационни метрики в настоящата глава са получени при чист протокол ($\text{adv_train} = \text{none}$); количествените ARI редове за адверсариално обучени варианти се разширяват в RED-MS.

3.4. Обяснимост

3.4.1. Grad-CAM

Grad-CAM [86] комбинира градиентите на целевия клас (Fake) с активациите на избран конволюционен слой за получаване на пространствена топлинна карта. Качествените примери показват, че активирането при коректно класифицирани fake изображения попада върху централни лицеви зони и текстурни преходи.



Фиг. 12. RED: локализация на доказателства чрез Grad-CAM.

3.4.2. Сравнение на методи за обяснимост

Сравняват се Grad-CAM, Layer-wise Relevance Propagation (LRP) [88], Integrated Gradients, LIME [89], SHAP, RISE, разгъване на вниманието (attention rollout) и AttnLRP. RED използва Grad-CAM като основен инструмент за CNN мрежите заради ниската цена и пряката интерпретируемост [90].

3.4.3. Обосновка и ограничения на метриците ARI и EF

ARI е дефинирана като отношение, а не абсолютна разлика, за да е сравнима между модели с различни изходни нива на точност [91]. EF (вярност на обясненията) формализира корелацията между картата на значимост и еталонна маска, но остава качествена в RED поради липсата на достатъчен брой пикселни маски, изисквайки пълен процес на оценка за количествено отчитане в RED-MS.

3.4.4. Обяснимост в RED рамката

Grad-CAM за Xception/ResNet-50 дава локализуеми обяснения, съгласувани с форензични очаквания; ViT клонът изисква отделни XAI методи, методологично дефинирани тук и емпирично разширени в RED-MS.

3.5. Експериментална методология

3.5.1. Набор от данни и разделения

Основните експерименти използват Kaggle real/fake faces набор със стратифицирано разделяне 70/15/15, преобразмеряване до 224×224 и ImageNet нормализация.

3.5.2. Сравнявани модели и софтуер

Сравняват се Xception, ResNet-50 и ViT-B/16 (Grad-CAM приложим само за първите две) върху стек PyTorch, timm, pytorch-grad-cam и scikit-learn.

3.5.3. Хиперпараметри на обучението

Споделена конфигурация: 10 епохи, batch size 32, learning rate 0.0003 [92], adv_train=none за референтния протокол, $\epsilon = 8$ (скала 0-255), $\alpha = 2$, PGD стъпки = 4, seed = 42 [93].

3.5.4. Метрики

Класификационните метрики включват Accuracy, Precision, Recall, F1 и AUC-ROC (двукласова); устойчивостта се отчита чрез ARI, а обяснимостта се оценява качествено чрез Grad-CAM.

3.5.5. Възпроизводимост и протокол

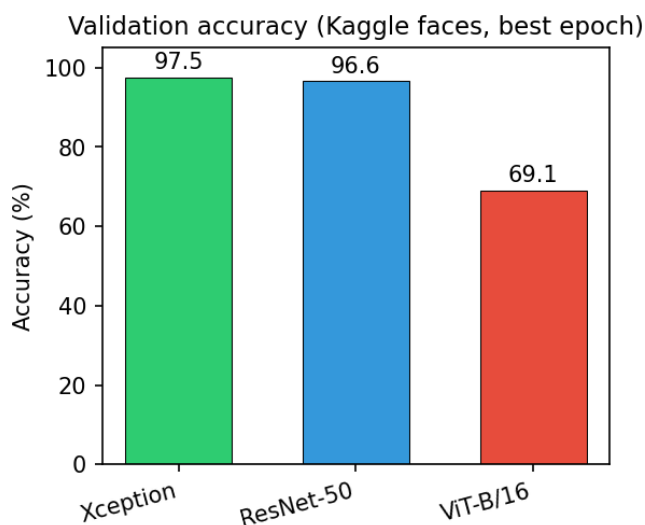
Възпроизводимостта се постига чрез фиксиран $\text{seed} = 42$, споделен протокол между трите основни мрежи и изцяло външна YAML/CLI конфигурация [94].

3.6. Експериментални резултати

3.6.1. Количествени резултати (валидация, чист протокол)

Таблица 3.1. Най-добри валидационни метрики по основна мрежа (Kaggle 70/15/15, $\text{adv_train}=\text{none}$)

Метрика	Xception	ResNet-50	ViT-B/16
Accuracy	97.53%	96.64%	69.08%
F1	97.51%	96.69%	67.69%
AUC-ROC	0.9977	0.9959	0.7521



Фиг. 22. Точност (%) по основна мрежа при най-добра валидационна епоха.

3.6.2. Качествена обяснимост

Панелите на Grad-CAM показват, че при правилно предсказани fake изображения активирането е около лицеви структури и потенциални зони на граница на смесване, докато при real изображения фокусът е върху естествени лицеви признаци.

3.6.3. Изпълнение на устойчивостта при RED

RED дефинира възпроизводимо FGSM/PGD адверсариално обучение и метрика ARI при L_∞ с пълен достъп; компромисът точност-устойчивост при $\text{adv_train} \in \{\text{fgsm}, \text{pgd}\}$ се пренася методологично в RED-MS.

3.7. Дискусия

Силната чиста производителност на CNN референтните подходи (Xception 97.53%, ResNet-50 96.64%) спрямо ViT-B/16 (69.08%) е специфична за протокола (10 епохи, ограничен бюджет) и не следва да се екстраполира като обща йерархия между архитектурите. Разделението между методологически дефинирано и емпирично доказано предотвратява надценяване, а обяснимостта чрез Grad-CAM е позиционирана като инструмент „с човек в цикъла“, без претенция за количествена оценка тук.

Резултатите от настоящата глава следва да се четат заедно с ограничението за модела на заплахата: устойчивостта, измерена чрез ARI, отразява единствено L_∞ -ограничени атаки при пълен достъп до градиентите на модела. Това е консервативен, но не изчерпателен сценарий – реален атакуващ без достъп до архитектурата би разполагал с по-малко информация, докато адаптивен атакуващ, специално проектиран срещу защитата на RED,

би представлявал по-строг тест. Затова докладваните стойности на ARI следва да се интерпретират като горна граница на устойчивостта при дадения бюджет ϵ , а не като гаранция за поведение при други модели на заплаха. Тази предпазливост в интерпретацията е последователна с методологическия принцип, прилаган в цялата дисертация: всяко количествено твърдение се придружава от изричен протокол и граници на валидност.

3.8. Ограничения

Обхватът на атаките е ограничен до L_∞ FGSM/PGD при пълен достъп ($\epsilon = 4/255$ и $8/255$); EF се анализира само качествено; резултатът на ViT (69.08%) е специфичен за набора и обучението; Grad-CAM не се прилага директно към ViT; експериментите са върху статични изображения, без претенция за реално време [96].

3.9. Изводи към трета глава

- RED дефинира възпроизводимо FGSM/PGD адверсариално обучение при L_∞ ($\epsilon = 4/255$ и $8/255$) и метриката $ARI = \text{Acc_adv}/\text{Acc_clean}$, която представлява формална рамка за оценка на компромиса точност-устойчивост (задача 4).
- При идентичен протокол конволюционните референтни подходи постигат силна чиста точност (Xception 97.53%, ResNet-50 96.64%) спрямо ViT-B/16 (69.08%); резултатът е специфичен за набора и бюджета на обучение.
- Разграничението между методологически дефинирано и емпирично доказано предотвратява надценяване: чистата точност задава горна граница преди стрес-теста, а не самата устойчивост.
- Обяснимостта чрез Grad-CAM е позиционирана като инструмент „с човек в цикъла“, който качествено потвърждава форензичната релевантност, но не замества количествената оценка (EF).
- Твърденията са ограничени до L_∞ атаки с пълен достъп; устойчивост срещу черна кутия и трансферни атаки не се претендира. Главата подготвя обединяването на адверсариалните и XAI компонентите с многопоточната архитектура в RED-MS.

ГЛАВА 4. МЕТОДОЛОГИЯ ЗА НАДЕЖДНО И ОБОСНОВАНО МНОГОПОТОЧНО ОТКРИВАНЕ НА ИИ-ГЕНЕРИРАНИ ЛИЦА

Главата завършва изследователската траектория чрез RED-MS (Robust and Explainable Detection – Multi Stream), унифицирана архитектура, която вгражда многопоточното три-класово откриване в адверсариалното обучение и обяснимостта на RED. Разделът изпълнява задача 5: методологичен синтез, експериментален протокол и сравнителен анализ.

4.1. Цел, експериментални етапи и обект на изследването

4.1.1. Мотивация

Точност, устойчивост и прозрачност често се разработват разделено. RED-MS реализира единен процес, в който се запазва три-класовата постановка, моделът се закалява срещу L_∞ -ограничени FGSM/PGD атаки, решенията се обосновават с локализация на доказателства, а обобщаването се измерва извън разпределението (OOD).

4.1.2. Цел и експериментални етапи

Целта е да се проектира, имплементира и дефинира протокол за емпирична оценка на RED-MS със сравнителен анализ спрямо хибридният метод, многопоточният референтен подход и RED. Експерименталните етапи са E1 (интеграция на потоците), E2 (FGSM/PGD адверсариално обучение), E3 (Grad-CAM и EF върху FF++ маски), E4 (разгъване на вниманието срещу Grad-CAM), E5 (zero-shot OOD оценка), E6 (унифицирана сравнителна таблица) [97].

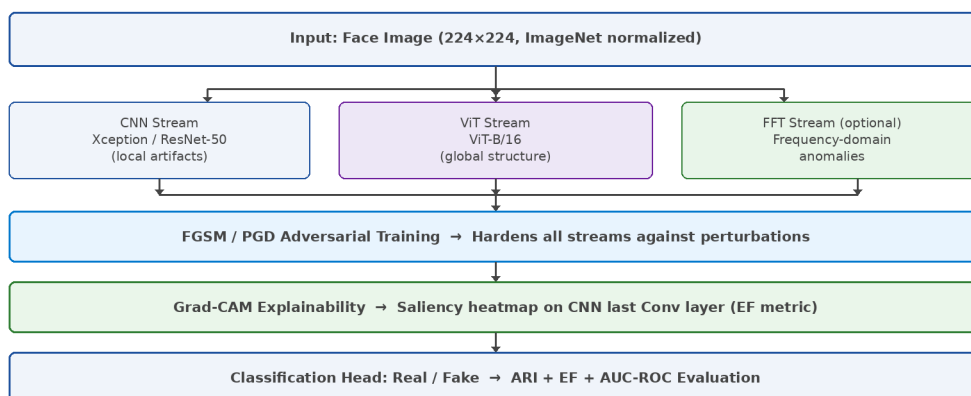
4.1.3. Обект и обхват

Обектът са статични изображения, а именно автентични, с размяна на лице и напълно ИИ-генерирани. В обхвата са L_∞ FGSM/PGD при white-box достъп, Grad-CAM върху Xception потока и стратифицирано разделяне 70/15/15. Извън обхвата остават видео-времева консистентност, атаки тип черна кутия и трансферни атаки [98], [99].

4.2. Методология на RED-MS

4.2.1. Общ преглед на процеса

Процесът следва модулен поток: предобработка (224×224 , ImageNet нормализация), три паралелни потока за признаци, сливане чрез внимание с динамични тегла α_v , α_c , α_f , адверсариално обучение чрез FGSM/PGD аугментация и последваща обяснимост чрез Grad-CAM и EF.



Фиг. 26. Концептуална архитектура RED-MS.

4.2.2. Многопоточно извличане на признаци

ViT-B/16 (timm, ImageNet предобучение) улавя глобална кохерентност; Xception [101] фокусира локални граници и компресионни артефакти и е съвместим с Grad-CAM; FFT потокът използва log-magnitude спектър след fftshift за дискриминация на спектралните отпечатъци [102].

4.2.3. Сливане чрез внимание и класификация

Сливаният вектор $h = \alpha_v \cdot e_{ViT} + \alpha_c \cdot e_{CNN} + \alpha_f \cdot e_{FFT}$ ($\sum \alpha_i = 1$) се подава на пълносвързан слой и softmax за три класа (real, face_swapped, ai_generated), оптимизиран чрез крос-ентропия $\mathcal{L}_{ce} = -\sum_c y_c \cdot \log(\hat{y}_c)$.

4.2.4. Адверсариално обучение

Конфигурацията red_ms_pgd използва $\epsilon = 8/255$, $\alpha = 2/255$ и 3 PGD стъпки при обучение, до 10 при оценка, с комбинирана загуба $\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{cons}$ и метрика $ARI = Acc_{adv} / Acc_{clean}$.

4.2.5. Обяснимост и вярност на обясненията

Grad-CAM се прилага само към Xception потока. При налична маска M и карта на значимост S: $EF_{Pearson} = corr(S, M)$, допълнено с IoU и Jaccard. EF се измерва за чисти и PGD-пертурбирани входове за тест на стабилността на обясненията.

4.2.6. Протокол за възпроизводимост на експериментите

Всяка стъпка E1-E9 оставя артефакти за независима проверка, а именно конфигурационен YAML файл с хеш сума, git commit, фиксирана начална стойност и контролна точка с най-добра епоха [103]. Верифицираното PGD изпълнение е извършено с align=false.

4.3. Етапи на експерименталната работа

Етапите E1-E9 обхващат: подготовка на данни (E1), обучение RED-MS clean (E2), обучение RED-MS PGD (E3), оценка в разпределението (E4), адверсариална оценка (E5), оценка на EF (E6), zero-shot OOD оценка (E7), Grad-CAM/ViT фигури и сравнителна таблица (E8), синтез с публикациите (E9) [104], [105].

4.3.1. Логика на експерименталните етапи

Етапите образуват зависима верига: E1 определя етикетите и разделянето, E2-E3 изолират ефекта на защитата, E4-E5 разделят класификацията в разпределението от адверсариалния стрес-тест, E6 е условен на наличието на FF++ анотации, E7 тества обобщаемостта, а E8-E9 синтезират резултатите.

4.3.2. Критерии за приемане на резултат

Протоколът изисква документирана най-добра епоха, матрица на обърквания без систематичен колапс на клас, ARI редове за FGSM и PGD при двата ϵ и EF CSV с $n \geq 100$ маскирани проби, тъй като единично високо число без адверсариални и XAI редове не е достатъчно.

4.4. Софтуерни технологии и възпроизводимост

Технологиите включват Python 3, PyTorch, timm (ViT-B/16, Xception), pytorch-Grad-CAM, Dlib, OpenCV, Hugging Face набори от данни, scikit-learn [106] и NumPy/Pandas.

4.4.1. Управление на версии и зависимости

Главните версии на PyTorch и timm са фиксирани в requirements.txt, тъй като адверсариалните атаки и Grad-CAM са чувствителни към autograd и имената на слоевете.

4.4.2. Логове, таблици и проследимост

CLI еталонният тест записва метрики в CSV/JSON с времеви маркер; таблиците за класификационна точност и ARI са получени от потвърденото PGD изпълнение, а таблиците за EF и OOD гар се дефинират от протокола.

4.5. Подбор и подготовка на данни

4.5.1. В разпределението (основна оценка)

AI-vs-Deepfake-vs-Real съдържа 9 999 балансираны изображения; стратифицирано разделяне 70/15/15, seed 42, без припокриване обучение/тест [112].

4.5.2. Обяснимост (EF)

FaceForensics++ предоставя маски на ниво пиксел за класически методи за размяна на лице; за количествен EF са необходими минимум 100 маскирани проби.

4.5.3. Контрол на качеството и етични филтри

Модулът за данни прилага дедупликация по перцептивен хеш (pHash), при който тестово изображение се маркира като near-duplicate при разстояние на Хеминг ≤ 5 бита спрямо обучаващото множество. Одитът потвърди липса на съществено изтичане между подмножествата.

4.5.4. Out-of-distribution (OOD) сурогати с JPEG и шум

Аугментациите JPEG и Гаусов шум симулират пост-обработка в социални мрежи, но не заменят генератори извън разпределение; систематичният OOD отчет изисква WildRF или Deepfake-Eval-2024 без донастройка (zero-shot протокол E7).

4.6. Резултати и сравнителен анализ

4.6.1. Референтни резултати

Сравнителният синтез използва публикуваните числа от предходните глави: хибриден 94.8% (WhichFaceIsReal), многопоточен 93.7% обща точност, RED/Xception 97.53% чиста точност и AUC 0.9977.

4.6.2. Резултати от обучение и оценка на RED-MS PGD

Верифицираното изпълнение (AdamW, $lr = 1 \cdot 10^{-4}$, класово претеглена крос-ентропия, PGD 3 стъпки при обучение, $\epsilon = 8/255$, fp16 autocast, batch size 24, 12 епохи) постига най-добра контролна точка на епоха 11 с валидационна точност 0.9913. Върху тестовото разделяне (1502 изображения) RED-MS постига чиста точност 0.9814, macro-F1 0.9813, MCC 0.9720 [113].

Таблица 4.2. Класови показатели на RED-MS върху тестовото разделяне (1502 изображения)

Клас	Precision	Recall	F1	AUC
ai_generated	0.9784	0.9784	0.9784	0.9991
face_swapped	0.9754	0.9675	0.9714	0.9983
real	0.9901	0.9980	0.9940	0.9999
macro	-	-	0.9813	0.9991

При white-box оценка с $\epsilon = 8/255$ ARI при FGSM е 0.9288, а при PGD-10 е 0.8440 (Таблица 4.4), което потвърждава, че PGD обучението запазва съществена устойчивост без катастрофична деградация.

4.6.3. Унифицирана сравнителна таблица

Таблица 4.5. Архитектурно сравнение на постигнатите резултати в различните методи.

Подход	Поток и	Класове	Асс (%)	Набор	Обяснимост	Адв. устойчивост
Само форензичен	1	2	77.1	WhichFaceIsReal	Не	Не
Само CNN (ResNet-18)	1	2	91.4	WhichFaceIsReal	Не	Не
Хибриден CNN и форензичен	2	2	94.8	WhichFaceIsReal	Не	Не
Многопоточен (ViT, CNN и FFT)	3	3	93.7	AI-vs-Deepfake-vs-Real	Качествена (Grad-CAM)	Не
RED / Xception	1 (CNN)	2	97.53	Kaggle real/fake	Grad-CAM	FGSM/PGD
RED / ResNet-50	1 (CNN)	2	96.64	Kaggle real/fake	Grad-CAM	FGSM/PGD
RED / ViT-B/16	1 (ViT)	2	69.08	Kaggle real/fake	Не	FGSM/PGD
RED-MS clean	3	3	99.13	AI-vs-Deepfake-vs-Real	Grad-CAM (CNN поток)	Не
RED-MS PGD	3	3	98.14	AI-vs-Deepfake-vs-Real	Grad-CAM (CNN поток)	FGSM/PGD (ARI PGD-10 =

Подход	Потоц и	Клас ове	Асс (%)	Набор	Обяснимос т	Адв. устойчиво ст
						0.844)

4.6.4. Разширение *out-of-distribution* (OOD) и обобщаемост

Протоколът за OOD оценка измерва разликата между точността в разпределението и точността при невидени генератори чрез zero-shot върху WildRF и Deepfake-Eval-2024, прилаган за RED-MS PGD и двукласовите RED модели. Систематичното отчитане е етап E7 от протокола.

4.6.5. Обяснимост при ViT – качествени насоки

Разгъването на вниманието при ViT-B/16 се съпоставя с Grad-CAM при Xception върху едни и същи изображения: при реалните лица вниманието се концентрира върху биологично кохерентни зони, при размяната върху границите на смесване, а при ИИ-генерираните върху глобалната текстура и спектралните аномалии.

4.6.6. Оценка на метриците за точност, устойчивост и обяснимост

Проследяването на точността от 77.1% през 94.8% и 93.7% до 97.53% при двукласовия RED показва архитектурна, а не транзитивна еволюция. RED-MS обединява многопоточното представяне, адверсариалната защита и обяснимостта в единна рамка: чиста точност 98.14% и измерима адверсариална устойчивост ($ARI = 0.844$ при PGD-10), като количественият EF и систематичната OOD оценка остават дефинирани по протокол.

4.7. Изводи към четвърта глава

- RED-MS интегрира многопоточната три-класова архитектура (ViT + CNN + FFT) с адверсариалното обучение и обяснимостта на RED в единен възпроизводим конвейер, обхващащ задачи 1-5 на дисертационния труд.
- Потвърденото PGD изпълнение дава първите емпирични резултати за обединената рамка: чиста точност 98.14%, macro-F1 0.9813, MCC 0.9720 и измерима адверсариална устойчивост (ARI при FGSM 0.9288 и при PGD-10 0.844).
- PGD обучението запазва профила на класовете с най-слаб клас „размяна на лице“ ($F1 = 0.9714$), което потвърждава, че защитата не разрушава три-класовата дискриминативна способност (хипотеза H6, методологично).

ЗАКЛЮЧЕНИЕ

Настоящият труд проследява еволюцията към доверимо откриване на ИИ-генерирани лица и разменени лица в четири глави: аналитичен преглед (Глава 1) и три емпирически и синтезиращи глави (Глави 2-4). По-долу се обобщават изследователската траектория, изпълнението на поставените задачи, ограниченията и приоритетните бъдещи насоки.

Обобщение на изследователския труд

Изследователската траектория следва логическа еволюция: от единичен хибриден поток към многопоточна три-класова архитектура, след това към адверсариално устойчива и обяснима двукласова рамка (RED), и накрая към обединена RED-MS рамка, която съчетава многопоточност, адверсариална защита и обясним изкуствен интелект в един процес. Хибридният двукласов модел постига 94.8% точност и $F1 = 0.942$, надминавайки конфигурациите само с CNN (91.4% / 0.904) и само с форензичен поток (77.1% / 0.755). Три-класовата многопоточна архитектура постига 93.7% обща точност с $F1$ по класове 0.97 (реално), 0.90 (разменено лице) и 0.94 (ИИ-генерирано), $MCC \approx 0.90$ и $AUC > 0.98$. Рамката RED постига при чист протокол Xception 97.53% точност, ResNet-50 96.64% и ViT-B/16 69.08%. За обединената рамка RED-MS е проведено верифицирано PGD

изпълнение с чиста точност 98.14%, macro-F1 = 0.9813, MCC = 0.9720 и ARI при PGD-10 = 0.844.

Изпълнение на поставените задачи

Шестте задачи на дисертационния труд са изпълнени. Задача 1 (анализ на състоянието и дефиниране на задачите) е изпълнена в Глава 1. Задачи 2 и 3 (хибриден двукласов модел и три-класова многопоточна архитектура) са емпирично изпълнени и апробирани в публикации. Задача 4 (рамка RED с FGSM/PGD, метрика ARI и Grad-CAM) е изпълнена методологично с верифициран чист референтен протокол. Задача 5 (синтез RED-MS, сравнителен анализ и протокол извън разпределение) е реализирана концептуално, методологично и емпирично чрез верифицираното PGD изпълнение. Задача 6 (формулиране на приносите, ограниченията, етичните и правните аспекти и перспективите) е изпълнена в настоящото заключение и в раздела за приносите.

Ограничения на изследването

- Всички експерименти са върху статични изображения; времевата съгласуваност на ниво видео, rPPG биометрията и мултимодалното откриване са извън обхвата.
- Адверсариалната оценка обхваща само L_∞ -ограничени FGSM и PGD при пълен достъп (white-box); трансферни, адаптивни и атаки тип черна кутия не се претендират като оценени.
- ViT-B/16 постига 69.08% точност при същото Kaggle разделяне — резултат, специфичен за набора и обучението, а не общ отказ на архитектурата.
- Обяснимостта (EF) е качествена в RED; количественият EF върху FaceForensics++ маски е дефиниран методологично за RED-MS.
- Демографският одит за безпристрастност при недостатъчно представени групи остава незавършен.

Бъдещи насоки за изследване

- Систематично OOD измерване между генератори и набори от данни (WildRF, Deepfake-Eval-2024, FaceForensics++).
- Откриване на ниво видео: времева съгласуваност и rPPG сигнали.
- Адаптивна и black-box устойчивост: оценка срещу трансферни атаки и обучение, отчитащо атакуващия.
- Количествен EF и стабилност на адверсариалните обяснения върху ≥ 100 маскирани изображения.
- XAI, съвместим с ViT: разгъване на вниманието (attention rollout) и AttnLRP.
- Одит за демографска безпристрастност и стратегии за смекчаване.
- Продължаващо обучение към нововъзникващи генеративни архитектури.

НАУЧНО-ПРИЛОЖНИ И ПРИЛОЖНИ ПРИНОСИ

Научни приноси:

– Предложен е хибриден двукласов модел за откриване на синтетични лица, който комбинира дълбоки признакови вектори от конволюционна невронна мрежа и форензичен признак за съотношение на бели пиксели и класификатор от тип случайни гори, с което се повишава точността спрямо самостоятелно използване на всеки източник на информация.

– Предложена е три-класова многопоточна архитектура с механизъм за сливане чрез внимание, използване на пространствени и честотни признаци и три-класова етикетна семантика за разграничаване на реални изображения, изображения с разменено

лице и напълно генерирани от изкуствен интелект лица като отделни форензични категории.

- Предложена е модулна рамка RED (Robust and Explainable Detection) за устойчиво и обяснимо откриване на изображения на лица генерирани с ИИ, която интегрира адверсариално обучение, метрика за оценка на адверсариалната устойчивост и методи за последваща обяснимост, осигуряващи едновременно повишена устойчивост срещу адверсариални атаки и проследимост на решенията на модела.

- Предложена е обединена архитектура RED-MS (Robust and Explainable Detection – Muslti Stream) за доверимо откриване на изображения на лица генерирани с ИИ с многопоточно три-класово откриване в адверсариалния и обяснимия процес на рамката RED, която съчетава в единен модел повишена точност, устойчивост срещу адверсариални атаки и обяснимост на решенията.

Научно-приложни приноси:

- Предложена е методология за унифициран сравнителен анализ на разработените подходи за откриване на синтетични лица – форензичен, хибриден, многопоточен, както и рамките RED и RED-MS, която позволява проследяване на приноса на всеки следващ подход спрямо предходните.

- Предложени са метрики за устойчивост и за обяснимост като оперативни индикатори, които позволяват количествена оценка на адверсариалната устойчивост и на коректността на обясненията на моделите.

Приложни приноси:

- Разработена е софтуерна имплементация с отворен код за реализиране на рамките RED и RED-MS, която осигурява практическа основа за възпроизвеждане и доразвиване на предложените модели и архитектури.

- Проведени са YAML-конфигурируеми еталонни тестове с команден интерфейс, на базата на които са извършени експериментални анализи за осигуряване на пълна проследимост на началните стойности, разделянето на данните и хиперпараметрите при обучение на моделите.

- Разработена е стандартизирана процедура за предобработка на лицеви изображения, осигуряваща съгласувано подаване на данните към моделите за откриване, и стратегия за модериране и верификация на съдържание в режим с участие на човек в процеса на вземане на решение.

- Реализирани са Grad-CAM визуализации като инструмент за форензичен преглед, които подпомагат експертната проверка на решенията на модела и

документирани ограничения и насоки за отговорно внедряване на разработените инструменти за откриване в практически условия.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

1. **G. Kotov**, O. Nakov, M. Lazarova and P. Nakov, „Hybrid CNN and Forensic Approach for Detecting AI-Generated Human Faces,“ Proc. of 60th International Scientific Conference on Information Technology (ICEST), 26–28 June 2025, Ohrid, North Macedonia. DOI: 10.1109/ICEST66328.2025.11098362 (индексирана в Scopus и WoS)
2. **G. Kotov**, O. Nakov, M. Lazarova, P. Nakov and G. Georgiev, „Face Swapping Attack Detection Using Multi-Stream Feature Embeddings,“ Proc. of 2025 13th International Scientific Conference on Computer Science (COMSCI), 13–15 Sept. 2025, Sozopol, Bulgaria. DOI: 10.1109/COMSCI67172.2025.11225271 (индексирана в Scopus)
3. **G. Kotov**, P. Nakov and O. Nakov, „Adversarial Robustness and Explainability in AI-Generated Face Detection,“ Proc. of 15th International Scientific Conference Engineering, Technology and Systems (TechSys 2026), 14–16 May 2026, Plovdiv, Bulgaria. (предстоящо публикуване в MDPI Engineering Proceedings и индексиране в Scopus и WoS)
4. **G. Kotov**, „Robust and Explainable Multi-Stream Detection of Face Swapping and AI-Generated Face Attacks,“ Proc. of 2026 14th International Scientific Conference Computer Science (COMSCI), 13–15 Sept. 2026, Sozopol, Bulgaria (приета за публикуване)

ИЗПОЛЗВАНА ЛИТЕРАТУРА

- [1] J. Boyle, “Study shows most of us can’t tell difference between human face and one created by AI,” The Sunday Post. Available: <https://www.sundaypost.com> (accessed 18 June 2025).
- [2] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, “GAN-Generated Faces Detection: A Survey and New Perspectives,” in ECAI 2023, Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 2533–2542, 2023.
- [3] “Deepfake Statistics 2023,” Home Security Heroes. Available: <https://www.homesecurityheroes.com/state-of-deepfakes> (accessed 18 June 2025).
- [4] “What are deepfakes, and how can you spot them?,” Sumsb. Available: <https://sumsub.com/blog/what-are-deepfakes> (accessed 18 June 2025).
- [5] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” Int’l Conf. Learning Representations (ICLR), 2015.
- [6] C. Szegedy, et al., “Intriguing properties of neural networks,” Int’l Conf. Learning Representations (ICLR), 2014.
- [7] K. Patil, S. Kale, J. Dhokey, and A. Gulhane, “Deepfake detection using biological features: A survey,” arXiv:2301.05819, 2023.
- [9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014.
- [10] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” Int’l Conf. Learning Representations (ICLR), 2016.
- [11] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” Int’l Conf. Learning Representations (ICLR), 2019.

- [14] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” *Int’l Conf. Learning Representations (ICLR)*, 2018.
- [15] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
- [16] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.
- [17] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, “Alias-free generative adversarial networks,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 852–863.
- [18] M. Groh, Z. Epstein, C. Firestone, and R. Picard, “Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds,” *arXiv:2105.06496*, 2021.
- [19] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.
- [20] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” *Proc. 38th Int’l Conf. Machine Learning (ICML)*, 2021, pp. 8162–8171.
- [21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *Int’l Conf. Learning Representations (ICLR)*, 2021.
- [22] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, “Diffusion Models in Vision: A Survey,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.
- [23] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” *Int’l Conf. Learning Representations (ICLR)*, 2014.
- [24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.
- [25] C. Saharia, et al., “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 19572–19586.
- [27] A. Ramesh, et al., “Zero-shot text-to-image generation,” *Proc. 38th Int’l Conf. Machine Learning (ICML)*, 2021, pp. 8821–8831.
- [28] G. Marcus, E. Davis, and S. Aaronson, “A very preliminary analysis of DALL-E 2,” *arXiv:2204.13807*, 2022.
- [29] “SynthID,” Google DeepMind. Available: <https://deepmind.google/science/synthid> (accessed 18 June 2025).
- [32] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2387–2395.
- [33] K. Liu, et al., “DeepFaceLab: Integrated, flexible and extensible face swapping framework,” *Pattern Recognition*, vol. 141, 2023.
- [35] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, “On the Detection of Digital Face Manipulation,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5781–5790.
- [36] S. Barrington, E. A. Cooper, and H. Farid, “People are poorly equipped to detect AI-powered voice clones,” *Scientific Reports*, vol. 15, no. 1, 2025.
- [37] A. Vaswani, et al., “Attention is all you need,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [40] N. Chandra, et al., “Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024,” *arXiv:2503.02857*, 2025.
- [41] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 86–103.
- [42] R. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7890–7899.
- [44] D. A. Cocomini, N. Messina, C. Gennaro, and F. Falchi, “Combining EfficientNet and Vision Transformers for video deepfake detection,” *Proc. Int’l Conf. Image Analysis and Processing (ICIAP)*, 2022, pp. 219–229.

- [45] L. Verdoliva, “Media forensics and deepfakes: An overview,” *IEEE J. Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.
- [46] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” *Proc. IEEE Int’l Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.
- [47] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5039–5049.
- [48] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *Proc. Int’l Conf. Machine Learning (ICML)*, vol. 97, 2019, pp. 6105–6114.
- [49] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized input sampling for explanation of black-box models,” *British Machine Vision Conference (BMVC)*, 2018.
- [50] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face X-ray for more general face forgery detection,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5001–5010.
- [51] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” *Proc. Int’l Conf. Machine Learning (ICML)*, 2017, pp. 3319–3328.
- [52] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [53] I. Skorokhodov, S. Tulyakov, and M. Elhoseiny, “StyleGAN-V: A Continuous Video Generator With the Price, Image Quality and Perks of StyleGAN2,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3626–3636.
- [54] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” *Proc. IEEE/CVF Int’l Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.
- [61] A. Dosovitskiy, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *Int’l Conf. Learning Representations (ICLR)*, 2021.
- [74] P. Jaccard, “The distribution of the flora in the alpine zone,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [75] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” *Proc. IEEE Symp. Security and Privacy (S&P)*, 2017, pp. 39–57.
- [77] G. Kotov, P. Nakov, and O. Nakov, “Adversarial Robustness and Explainability in AI-Generated Face Detection,” *Proc. 15th Int’l Scientific Conference TechSys 2026 - Engineering, Technology and Systems*, Plovdiv, Bulgaria, 14–16 May 2026.
- [82] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Int’l Conf. Learning Representations (ICLR)*, 2015.
- [83] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” *Proc. 35th Int’l Conf. Machine Learning (ICML)*, 2018, pp. 274–283.
- [84] A. Paszke, et al., “PyTorch: An imperative style, high-performance deep learning library,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.
- [85] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, 2018, pp. 839–847.
- [86] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.
- [87] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.
- [88] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PLoS ONE*, vol. 10, no. 7, e0130140, 2015.
- [89] D. Hajialigol, H. Liu, and X. Wang, “XAI-CLASS: Explanation-enhanced text classification with extremely weak supervision,” *arXiv:2311.00189*, 2023.
- [90] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *Proc. IEEE Int’l Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

- [91] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [92] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *Int’l Conf. Learning Representations (ICLR)*, 2018.
- [93] J. Cohen, E. Rosenfeld, and J. Z. Kolter, “Certified adversarial robustness via randomized smoothing,” *Proc. Int’l Conf. Machine Learning (ICML)*, 2019, pp. 1310–1320.
- [94] R. Wightman, “PyTorch Image Models (timm),” GitHub, 2019. Available: <https://github.com/huggingface/pytorch-image-models>.
- [95] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.
- [96] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” *Int’l Conf. Learning Representations (ICLR) Workshop*, 2017.
- [97] Y. Li and S. Lyu, “Exposing DeepFake videos by detecting face warping artifacts,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.
- [98] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer, “Adversarial threats to DeepFake detection: A practical perspective,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 923–932.
- [99] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2185–2194.
- [100] G. Kotov, O. Nakov, and P. Nakov, “Robust and Explainable Multi-Stream Detection of Face Swapping and AI-Generated Face Attacks (RED-MS),” *2026 Int’l Scientific Conference on Computer Science (ComSci 2026)*, Dept. of Computer Systems, Technical University of Sofia, Bulgaria – accepted for publication.
- [101] I. Castillo Camacho, “Initialization methods of convolutional neural networks for detection of image manipulations,” *Signal and Image Processing*, Université Grenoble Alpes, 2021.
- [102] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 8780–8794.
- [103] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging frequency analysis for deep fake image recognition,” *Proc. 37th Int’l Conf. Machine Learning (ICML)*, 2020, pp. 3247–3258.
- [104] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [105] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.
- [107] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Two-stream neural networks for tampered face detection,” *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1831–1839.
- [108] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4190–4197.
- [109] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The DeepFake Detection Challenge (DFDC) dataset,” *arXiv:2006.07397*, 2020.
- [110] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *Proc. 32nd Int’l Conf. Machine Learning (ICML)*, 2015, pp. 448–456.
- [111] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *Int’l Conf. Learning Representations (ICLR)*, 2019.
- [112] N. Carlini and H. Farid, “Evading deepfake-image detectors with white- and black-box attacks,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 658–659.
- [113] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *Int’l Conf. Learning Representations (ICLR)*, 2015.
- [114] S. Ahmed, “Examining public perception and cognitive biases in the presumed influence of deepfakes threat,” *Asian J. Communication*, vol. 33, no. 3, 2023, pp. 308–331.
- [115] L. Lin, S. Santosh, M. Wu, X. Wang, and S. Hu, “AI-Face: A million-scale demographically annotated AI-generated face dataset and fairness benchmark,” *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 3503–3515.

- [116] An Overview into InterfaceGAN: Edit Facial Attributes of People Using GANs, Codex, Medium. Available: <https://medium.com/codex/an-overview-into-interfacegan-edit-facial-attributes-of-people-using-gans-34f2273d5941> (accessed 15 June 2026).
- [117] J. West and C. Bergstrom, WhichFaceIsReal, University of Washington. Available: <https://whichfaceisreal.com> (accessed 15 June 2026).
- [118] New Deepfake Method: HiFiFace, Unite.AI. Available: <https://www.unite.ai/new-deepfake-method-hififace/> (accessed 15 June 2026).
- [119] Deepfakes Creating New Challenges for Society, AP News. Available: <https://apnews.com/article/ap-top-news-artificial-intelligence-social-platforms-think-tanks-politics-bc2f19097a4c4fffaa00de6770b8a60d> (accessed 15 June 2026).
- [120] Computer Scientists Demonstrate the Potential for Faking Video, IJPR, 2017. Available: <https://www.ijpr.org/2017-07-21/computer-scientists-demonstrate-the-potential-for-faking-video> (accessed 15 June 2026).
- [121] Deepfake Statistics and AI-Generated Media, Statista, 2024. Available: <https://www.statista.com> (accessed 15 June 2026).
- [122] X. Yang, G. Su, J. Chen, N. Su and X. Ren, "Large Scale Identity Deduplication Using Face Recognition Based on Facial Feature Points," in: Biometric Recognition, CCBR 2011, Lecture Notes in Computer Science, vol. 7098, pp. 25-32, Springer, Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-25449-9_4
- [123] Y. Jeong, D. Kim, Y. Ro, P. Kim and J. Choi, "FingerprintNet: Synthesized Fingerprints for Generated Image Detection," in Computer Vision - ECCV 2022, Lecture Notes in Computer Science, vol. 13674, Springer, Cham, 2022. https://doi.org/10.1007/978-3-031-19781-9_5.

SUMMARY

METHODS FOR DISTINGUISHING AND RECOGNISING AI-GENERATED HUMAN FACE IMAGES FROM REAL HUMAN FACE IMAGES

Georgi Yanev Kotov

The widespread availability of generative artificial intelligence tools has fundamentally undermined trust in visual evidence. Synthetic human face images, produced by Generative Adversarial Networks (GAN), latent diffusion models such as Stable Diffusion, Imagen, and DALL·E, or by face-swapping pipelines, pose an operational risk to digital identity, biometric authentication, media verification, and law enforcement. A key forensic distinction motivating this dissertation is the differentiation between face-swapped images, which retain local blending boundary artefacts on an existing frame, and fully AI-generated portraits, which may appear globally coherent yet carry statistical anomalies in texture, symmetry, and spectral signatures. Detectors that collapse both manipulation types into a single class lose forensic information that is critical for moderation and investigation.

The goal of the dissertation is to design and evaluate methods for detecting AI-generated and identity-swapped faces with high accuracy within a three-class forensic taxonomy, adversarial robustness under L-infinity white-box constraints, post-hoc explainability, and unification in the RED-MS framework (Robust and Explainable Detection – Multi-Stream). Six research tasks are defined and fulfilled, progressing from analysis of the state of the art through design of a two-class hybrid model, a three-class multi-stream architecture, an adversarially hardened explainable framework (RED), and a unified RED-MS pipeline. Scientific methods include comparative literature analysis, neural architecture design with PyTorch, stratified data splitting (70/15/15 and 80/20), L-infinity-bounded adversarial evaluation with FGSM and PGD at $\epsilon = 4/255$ and $8/255$, and post-hoc explainability via Grad-CAM.

A hybrid two-class detector is designed combining a ResNet-18 semantic feature vector with a forensic white-pixel ratio descriptor and a Random Forest classifier demonstrating that semantic and pixel-level signals are complementary for detecting GAN-generated faces. A three-class multi-stream architecture is proposed unifying a Vision Transformer (ViT-B/16) for global dependencies, a CNN for local blending artefacts, and an FFT stream for spectral anomalies, fused via an attention mechanism. A RED (Robust and Explainable Detection) framework is suggested that integrates FGSM/PGD adversarial training, Adversarial Robustness Index ($ARI = \text{Acc}_{adv}/\text{Acc}_{clean}$), and Grad-CAM post-hoc explainability. A RED-MS (Robust and Explainable Detection – Multi-Stream) pipeline is proposed that unifies the multi-stream three-class detector with adversarial hardening and XAI in a single YAML-driven nine-stage pipeline.

The scope of claims is explicitly bounded to static images only, with no video-temporal or multimodal detection; L-infinity FGSM/PGD white-box attacks only, with no black-box, transfer, or adaptive attacks; and qualitative explainability fidelity only. Future directions include systematic out-of-distribution evaluation on WildRF and Deepfake-Eval-2024, video-level detection with rPPG signals, black-box and adaptive robustness, ViT-compatible XAI via attention rollout, demographic fairness auditing, and continual learning toward emerging generative architectures.