



## РЕЦЕНЗИЯ

върху дисертационен труд за придобиване на образователна и научна степен „Доктор“ в професионално направление 5.3 „Комуникационна и компютърна техника“ по научна специалност „Автоматизирани системи за обработка на информация и управление“

Автор на дисертационния труд: **маг. инж. Георги Янев Котов**

Тема на дисертационния труд: **Методи за разграничаване и разпознаване на изображения на човешки лица, генерирани с изкуствен интелект, от изображения на реални човешки лица**

Рецензент: **проф. д-р инж. Румен Иванов Трифонов**

### 1. Актуалност на дисертационния труд

Тематиката на дисертационния труд на маг. инж. Георги Котов е изключително актуална към 2026 година и се намира в центъра на един от най-динамично развиващите се и социално значими проблеми в съвременната компютърна наука. Авторът коректно идентифицира три взаимно усилващи се процеса, обуславящи актуалността: демократизацията на генеративните инструменти, масовото разпространение на манипулирано съдържание в социалните мрежи и нарастващите регулаторни изисквания за прозрачност и отчетност при автоматизирана модерация.

Съществуващите подходи — едноточни, хибридни, многопоточни, честотни и трансформерни — се разработват предимно изолирано, без интегрирана рамка, обединяваща точност, устойчивост на атаки и обяснимост в единна верифицирана система. Тази изследователска празнина е реална и нейното запълване представлява научно значим принос.

На криминалистично и правно равнище използването на синтетични изображения в наказателни производства, при верификация на самоличност и при журналистически разследвания поставя остра нужда от методи, които не само класифицират достоверно, но и обясняват решението си по начин, пригоден за съдебна употреба. Именно тук интеграцията на компонент за обяснимост (XAI) в предложената RED рамка е особено актуална и научно обоснована.

Авторът коректно определя и границите на своите твърдения, като не представя разработките като универсално решение на проблема.

### 2. Степен на познаване състоянието на проблема и творческа интерпретация на литературния материал

Авторът на дисертационния труд демонстрира много добро познаване на предметната област и основните съвременни направления в изследването на методи за откриване на ИИ-генерирани и манипулирани изображения на човешки лица.

Направен е последователен анализ на развитието на генеративните технологии — от ранните DCGAN и ProGAN, през StyleGAN семейството, до съвременните дифузионни модели и face-swap технологии. Криминалистично прецизното разграничение между глобално кохерентни ИИ-генерирани лица и face-swap изображения с локални артефакти е в основата на три-класовата таксономия и

свидетелства за реално разбиране на проблема, а не за повърхностно боравене с понятията.

Таксономията на подходите за откриване е изградена с изследователска зрялост. Авторът коректно разграничава едноточните CNN методи (MesoNet, ResNet, Xception, Face X-ray) от хибридните подходи, комбиниращи семантични и криминалистични признаци, и от многопоточните архитектури с паралелно кодиране чрез ViT, CNN и FFT с attention-базирано сливане.

Добро познаване на областта е демонстрирано и при анализа на състезателните атаки и устойчивостта на моделите. Авторът разглежда FGSM и PGD атаките и поставя въпроса за устойчивостта като съществено условие за практическата приложимост на системите за криминалистично откриване, като заедно с това се отчита необходимостта от обяснимост на решенията, като XAI и Grad-CAM са разгледани като средства за подпомагане на интерпретацията и експертния анализ.

Дефинирани са и са използвани показатели за класификационна ефективност, устойчивост на атаки и обяснимост, като чрез тях се изгражда последователна рамка за сравнителна оценка на разработените подходи. Това показва разбиране не само на алгоритмичната страна на проблема, но и на необходимостта резултатите да бъдат оценявани чрез ясно дефинирани експериментални протоколи.

Вземайки предвид всичко това може да се заключи, че дисертантът притежава необходимата теоретична и методологична подготовка за провеждане на самостоятелно научно изследване в разглежданата област.

### **3. Съответствие на предложената методика на изследване и поставените цел и задачи на дисертационния труд**

Целта на дисертационния труд е ясно формулирана – проектиране и оценяване на методи за откриване на ИИ-генерирани и лица с разменена идентичност, като се отчитат едновременно точността, трикласовата класификация, устойчивост на атаки и обяснимостта.

Предложената методика е логически и последователно изградена. Налице е ясно проследима изследователска траектория: първоначално се изследва хибриден двукласов подход, след това се преминава към трикласова многопоточна архитектура, въвеждат се състезателно обучение и XAI, а накрая отделните направления се интегрират в рамката RED-MS. Поставените задачи са пряко свързани с отделните етапи на изследването.

Съответствието между цел, задачи, съдържание и резултати е отлично. Трудът следва строга логика на контролирано усложняване и всички ключови твърдения са подкрепени с конкретни експериментални данни.

### **4. Характеристика на естеството и оценка на достоверността на материала, върху който се градят приносите на дисертационния труд**

Изследванията в дисертационния труд се базират върху три вида материали: научна литература, еталонни набори от данни и експериментални резултати от собствени изследвания.

Разгледани са 123 литературни източника от утвърдени бази данни — IEEE Xplore, ACM Digital Library, arXiv, като основно са разгледани публикации след 2020 г. Включени са фундаментални трудове по GAN, дифузионни модели и състезателно обучение, както и специализирани изследвания по детекция на синтетични лица.

Съществена характеристика на дисертационния труд е неговият експериментален и сравнителен характер. Разработените подходи са оценявани чрез ясно дефинирани показатели за класификационна ефективност, сред които Accuracy, Precision, Recall и F1-score, а в последващите изследвания са включени показатели за устойчивост на атаки и обяснимост. При хибридният модел е извършено сравнение между CNN-only, Forensic-only и Hybrid конфигурации при еднакъв експериментален протокол, което позволява да се оцени конкретният ефект от комбинирането на семантични и криминалистични признаци. Получените резултати – 91.4% точност за CNN-only, 77.1% за Forensic-only и 94.8% за Hybrid.

Достоверността на експерименталните изследвания се подкрепят от използването на независими тестови множества, разделяния на пластове на данните, фиксирана начална стойност, документираните хиперпараметри и проследим експериментален процес. Това създава необходимите предпоставки за възпроизводимост на получените резултати и за коректно сравнение между отделните модели. Допълнителна положителна страна е разработването на YAML-конфигурируеми тестове и команден интерфейс, чрез които се проследяват началните стойности, разделянето на данните и параметрите на експериментите.

Особено внимание заслужава проверката на устойчивост на атаки. За RED-MS е посочено верифицирано изпълнение на PGD експеримента, при което е отчетена 98.14% точност върху чисти данни и ARI = 0.844 при PGD-10. Тези резултати представляват количествена експериментална основа за направените изводи относно устойчивостта на предложената архитектура.

Ограничение в изследванията, което може да окаже влияние върху достоверността е фактът, че в различните етапи на изследването използват различни набори от данни, което затруднява прякото сравнение на абсолютните стойности на точността.

## **5. Приноси на дисертационния труд**

Приносите са ясно формулирани, логично структурирани и съответстват на поставената цел и изпълнените задачи. Те могат да бъдат разгледани в три групи – научни, научно-приложни и приложни.

### *- Научни приноси*

Най-същественият научен принос е последователното изграждане на методология за доверимо откриване на синтетични лица, която интегрира три-класова криминалистична таксономия, многопоточно извличане на признаци, устойчивост на атаки и обяснимост. Предложеният хибриден двукласов модел (CNN + криминалистичен признак) демонстрира емпирично ползата от комбиниране на семантични и пикселни сигнали. Три-класовата многопоточна архитектура със сливане чрез проверка успешно разграничава реални изображения, изображения с разменено лице и напълно ИИ-генерирани лица като отделни категории, което е важно форензично уточнение спрямо класическото бинарно „реално/фалшиво“. Рамката RED формализира състезателното обучение (FGSM/PGD), въвежда метриката ARI и интегрира Grad-CAM. Най-значимият научен резултат е обединената архитектура RED-MS, която съчетава многопоточността с защита от атаки и обяснимостта в единен процес и постига високи емпирични показатели (98.14 % точност и ARI = 0.844 при PGD-10).

### *- Научно-приложни приноси*

Предложената методология за унифициран сравнителен анализ позволява проследяване на приноса на всеки следващ етап спрямо предходните.

Въведените метрики ARI и EF имат практическа стойност като оперативни индикатори за оценка на устойчивост и вярност на обясненията.

- *Приложни приноси*

Разработената софтуерна реализация с отворен код, YAML-конфигурируеми експерименти и стандартизирана предобработка осигуряват висока възпроизводимост. Grad-CAM визуализациите и формулираните насоки за отговорно използване повишават практическата приложимост на моделите в режим „човек в цикъла“.

Приносите са оригинални, добре обосновани и с ясно практическо значение. Ограниченията (обхват само върху статични изображения, white-box атаки, предимно качествена обяснимост) са коректно декларирани от автора и не намаляват съществено стойността на постигнатото.

Приемам приносите и считам, че те напълно съответстват на изискванията за присъждане на образователната и научна степен „доктор“.

## **6. Степен на личното участие на дисертанта в приносите**

Структурата, логиката и вътрешната последователност на труда носят белезите на авторска концептуална работа — методологичният избор на контролирано усложняване, последователното изграждане на хибридният модел, многопоточната архитектура, RED и RED-MS рамката, както и въведените оригинални метрики ARI и EF, са взаимно обвързани и свидетелстват за единна изследователска позиция.

Не съм установил некоректно заимстване и смятам, че цитирането е систематично, използването на чужди резултати е последователно, а собствените приноси са ясно разграничени от съществуващите подходи в литературата.

Експерименталната работа е реализирана в конкретна изчислителна среда по проект IDEAS с имплементация в PyTorch, конфигурируем процес и фиксирани seed стойности — детайли, характерни за самостоятелно проведено изследване.

## **7. Преценка на публикациите по дисертационния труд**

Маг. инж. Георги Котов е направил 4 публикации, свързани с основните приноси на дисертацията, които са реферирани в SCOPUS. Една от тях е самостоятелна, а останалите три публикации са изготвени в съавторство с научните ръководители на докторанта. Те покриват последователно хибридният модел, многопоточната архитектура, рамката RED и синтеза RED-MS, което свидетелства за систематична апробация на резултатите.

С посочените публикации се покрива изискуемия минимум, заложен в Правилника за условията и реда за придобиване на образователна и научната степен "Доктор" на ТУ-София.

## **8. Използване на резултатите от дисертационния труд в научната и социална практика**

Резултатите от дисертационния труд могат да намерят приложение в научните изследвания по медийна криминалистика и откриване на синтетично съдържание, както и в практически системи за верификация на изображения с участие на човек в процеса на вземане на решение.

## **9. Съответствие на автореферата с изискванията за изготвянето му и адекватност на отразяване на основните положения и приносите на дисертационния труд**

Докторантът е изготвил автореферат в обем от 33 печатни страници на български език и 30 страници на английски език. Авторефератът представя в резюме съдържанието на дисертационния труд, като поставя ударението върху негови основни елементи и представя точно приносите, постигнати в него.

Документът е съобразен изцяло с установените изисквания на ТУ — София за оформяне на автореферат.

## **10. Мнения, препоръки и бележки**

Дисертационният труд на маг. инж. Георги Котов представлява завършена, методологично издържана и технически задълбочена научна разработка по изключително актуална тема. Приносите са оригинални, експериментално обосновани и съответстват на поставената цел. Авторът демонстрира добра теоретична подготовка, способност за интегриране на разнородни технически подходи в единна концептуална рамка и коректно декларира ограниченията на изследването.

Позволявам си да направя следната препоръка за бъдещи изследвания:

- Да се завърши количествената оценка на обяснимостта (EF) върху маски от FaceForensics++.

Като забележка бих посочил, че сравненията между различните етапи следва да се интерпретират внимателно, тъй като са получени върху различни набори от данни.

## **ЗАКЛЮЧЕНИЕ**

Дисертационният труд напълно отговаря на изискванията на Закона за развитие на академичния състав на Република България, на Правилника за неговото приложение и на Правилника за условията и реда за придобиване на научни степени в Техническия университет — София.

Докторантът е извършил добре организирана изследователска работа, стигнал е до правилни изводи и въз основа на това е направил разработки с научен и приложен характер.

Въз основа на това, препоръчвам на уважаемото научно жури да присъди на маг. инж. Георги Янев Котов образователната и научна степен „доктор“ в професионалното направление 5.3 „Комуникационна и компютърна техника“ и научната специалност „Автоматизирани системи за обработка на информация и управление“.

02.09.2026 г.

София,

**РЕЦЕНЗЕНТ:**

*/проф. д-р инж. Р. Трифонов/*

# REVIEW

on the dissertation for the acquisition of the educational and scientific degree "**Doctor**" in the professional field 5.3 "**Communication and Computer Engineering**" in the scientific specialty "**Automated Information Processing and Management Systems**"

Author of the dissertation: **M.Eng. Georgi Yanev Kotov**

Topic of the dissertation: **Methods for distinguishing and recognizing images of human faces generated with artificial intelligence from images of real human faces**

Reviewer: Prof. PhD Eng. Roumen Ivanov Trifonov

## 1. Relevance of the dissertation work

The topic of the dissertation work of M.Sc. Eng. Georgi Kotov is extremely relevant by 2026 and is at the center of one of the most dynamically developing and socially significant problems in modern computer science. The author correctly identifies three mutually reinforcing processes that determine the relevance: the democratization of generative tools, the mass distribution of manipulated content in social networks, and the increasing regulatory requirements for transparency and accountability in automated moderation.

Existing approaches — single-stream, hybrid, multi-stream, frequency and transformer — are developed mainly in isolation, without an integrated framework that unites accuracy, resistance to attacks and explainability in a single verified system. This research gap is real and filling it represents a scientifically significant contribution.

At the forensic and legal level, the use of synthetic images in criminal proceedings, identity verification and journalistic investigations creates an urgent need for methods that not only classify reliably, but also explain their decisions in a manner suitable for judicial use. It is here that the integration of an explainability component (XAI) into the proposed RED framework is particularly relevant and scientifically sound.

The author correctly defines the limits of his claims, not presenting the developments as a universal solution to the problem.

## 2. Degree of familiarity with the state of the problem and creative interpretation of the literature

The author of the dissertation demonstrates very good knowledge of the subject area and the main contemporary directions in the research on methods for detecting AI-generated and manipulated images of human faces.

A consistent analysis of the development of generative technologies has been carried out — from the early DCGAN and ProGAN, through the StyleGAN family, to contemporary diffusion models and face-swap technologies. The forensically precise distinction between globally coherent AI-generated faces and face-swap images with local artefacts forms the basis of the three-class taxonomy and attests to a genuine understanding of the problem, rather than a superficial handling of the concepts.

The taxonomy of detection approaches is constructed with research maturity. The author correctly distinguishes single-stream CNN methods (MesoNet, ResNet, Xception, Face X-ray) from hybrid approaches that combine semantic and forensic features, and from multi-stream architectures with parallel encoding via ViT, CNN and FFT with attention-based fusion.

A good knowledge of the field is also demonstrated in the analysis of adversarial attacks and model robustness. The author examines FGSM and PGD attacks and poses the question of robustness as an essential condition for the practical applicability of forensic detection systems, while at the same time taking into account the need for explainability of decisions, with XAI and Grad-CAM considered as means of supporting interpretation and expert analysis.

Indicators for classification efficiency, robustness to attacks and explainability have been defined and used, and through them a consistent framework for comparative evaluation of the developed approaches is built. This shows an understanding not only of the algorithmic side of the problem, but also of the necessity that results be assessed through clearly defined experimental protocols.

Taking all of the above into account, it may be concluded that the doctoral candidate possesses the necessary theoretical and methodological preparation for conducting independent scientific research in the field under consideration.

### **3. Compliance of the proposed research methodology with the set goals and objectives of the dissertation**

The goal of the dissertation is clearly formulated – design and evaluation of methods for detecting AI-generated and swapped identities, taking into account both accuracy, three-class classification, resistance to attacks and explainability.

The proposed methodology is logically and consistently built. There is a clearly traceable research trajectory: initially a hybrid two-class approach is investigated, then a three-class multi-threaded architecture is switched, adversarial learning and XAI are introduced, and finally the individual directions are integrated into the RED-MS framework. The tasks set are directly related to the individual stages of the research.

The correspondence between the goal, objectives, content and results is excellent. The work follows a strict logic of controlled complication and all key statements are supported by specific experimental data.

### **4. Characterization of the nature and assessment of the reliability of the material on which the contributions of the dissertation are based**

The research in the dissertation is based on three types of materials: scientific literature, reference data sets and experimental results from own research.

A total of 123 literary sources from established databases — IEEE Xplore, ACM Digital Library, and arXiv — have been reviewed, with the primary focus on publications after 2020. Fundamental works on GANs, diffusion models and adversarial training are included, as well as specialized studies on the detection of synthetic faces.

An essential characteristic of the dissertation is its experimental and comparative nature. The developed approaches are evaluated through clearly defined indicators of classification efficiency, including Accuracy, Precision, Recall and F1-score, and in subsequent studies, indicators of resistance to attacks and explainability are included. In the hybrid model, a comparison was made between CNN-only, Forensic-only and Hybrid configurations under the same experimental protocol, which allows to assess the specific effect of combining semantic and forensic features. The results obtained – 91.4% accuracy for CNN-only, 77.1% for Forensic-only and 94.8% for Hybrid.

The reliability of the experimental studies is supported by the use of independent test sets, data layer separations, a fixed initial value, documented hyperparameters and a traceable experimental process. This creates the necessary prerequisites for the reproducibility of the obtained results and for a correct comparison between the individual models. An additional positive side is the development of YAML-configurable

tests and a command interface, through which the initial values, data separation and parameters of the experiments are tracked.

The verification of resistance to attacks deserves special attention. A verified implementation of the PGD experiment is reported for RED-MS, in which 98.14% accuracy on clean data and  $ARI = 0.844$  for PGD-10 were reported. These results provide a quantitative experimental basis for the conclusions drawn regarding the robustness of the proposed architecture.

A limitation of the studies, which may affect the reliability, is the fact that different datasets are used in different stages of the study, which makes it difficult to directly compare absolute accuracy values.

## **5. Contributions of the dissertation**

The contributions are clearly formulated, logically structured and correspond to the set goal and the tasks performed. They can be considered in three groups – scientific, scientific-applied and applied.

### *- Scientific contributions*

The most significant scientific contribution is the consistent construction of a methodology for reliable synthetic face detection, which integrates a three-class forensic taxonomy, multi-threaded feature extraction, attack resistance and explainability. The proposed hybrid two-class model (CNN + forensic feature) empirically demonstrates the benefit of combining semantic and pixel signals. The three-class multi-threaded architecture with fusion by verification successfully distinguishes real images, face-swapped images and fully AI-generated faces as separate categories, which is an important forensic refinement compared to the classic binary “real/fake”. The RED framework formalizes adversarial learning (FGSM/PGD), introduces the ARI metric and integrates Grad-CAM. The most significant scientific result is the unified RED-MS architecture, which combines multithreading with attack protection and explainability in a single process and achieves high empirical indicators (98.14% accuracy and  $ARI = 0.844$  at PGD-10).

### *- Scientific and applied contributions*

The proposed methodology for unified comparative analysis allows tracking the contribution of each subsequent stage compared to the previous ones. The introduced metrics ARI and EF have practical value as operational indicators for assessing the robustness and fidelity of explanations.

### *- Applied contributions*

The developed open-source software implementation, YAML-configurable experiments and standardized preprocessing ensure high reproducibility. Grad-CAM visualizations and formulated guidelines for responsible use increase the practical applicability of the models in the “man in the loop” mode.

The contributions are original, well-founded and of clear practical significance. The limitations (scope only on static images, white-box attacks, mainly qualitative explainability) are correctly declared by the author and do not significantly reduce the value of the achievement.

I accept the contributions and believe that they fully comply with the requirements for awarding the educational and scientific degree “doctor”.

## **6. Degree of the doctoral candidate’s personal involvement in the contributions**

The structure, logic and internal coherence of the work bear the marks of the author's own conceptual effort — the methodological choice of controlled progressive complexity, the consistent construction of the hybrid model, the multi-stream architecture, the RED

and RED-MS framework, as well as the introduced original metrics ARI and EF, are mutually interrelated and attest to a unified research position.

I have not identified improper borrowing and consider that the citation is systematic, the use of others' results is consistent, and the candidate's own contributions are clearly distinguished from existing approaches in the literature.

The experimental work has been carried out in a specific computational environment under the IDEAS project, with an implementation in PyTorch, a configurable process and fixed seed values — details characteristic of independently conducted research.

## **7. Assessment of the publications related to the dissertation**

M.Eng. Georgi Kotov has produced 4 publications related to the main contributions of the dissertation, which are indexed in SCOPUS. One of them is single-authored, while the remaining three have been prepared in co-authorship with the doctoral candidate's supervisors. They cover in sequence the hybrid model, the multi-stream architecture, the RED framework and the RED-MS synthesis, which attests to a systematic validation of the results.

With the stated publications, the required minimum set out in the Regulations on the terms and procedure for acquiring the educational and scientific degree "Doctor" of the Technical University of Sofia is fulfilled.

## **8. Use of the results of the dissertation in scientific and social practice**

The results of the dissertation can find application in scientific research on media forensics and the detection of synthetic content, as well as in practical systems for image verification with a human in the decision-making loop.

## **9. Compliance of the abstract with the requirements for its preparation and adequacy of reflection of the main points and contributions of the dissertation work**

The doctoral candidate has prepared an abstract in the volume of 33 printed pages in Bulgarian and 30 pages in English. The abstract presents in summary the content of the dissertation work, emphasizing its main elements and accurately presenting the contributions achieved in it.

The document is fully consistent with the established requirements of TU - Sofia for the formation of an abstract.

## **10. Opinions, recommendations and notes**

The dissertation work of M. Eng. Georgi Kotov represents a complete, methodologically sound and technically in-depth scientific work on an extremely topical topic. The contributions are original, experimentally substantiated and correspond to the set goal. The author demonstrates good theoretical preparation, the ability to integrate diverse technical approaches into a single conceptual framework and correctly declares the limitations of the study.

I would like to make the following recommendation for future research:

- To complete the quantitative assessment of explainability (EF) on masks from FaceForensics++.

As a note, I would like to point out that comparisons between the different stages should be interpreted with caution, since they were obtained on different data sets.

## **CONCLUSION**

The dissertation fully meets the requirements of the Law on the Development of the Academic Staff of the Republic of Bulgaria, the Regulations for its implementation and the Regulations on the conditions and procedure for acquiring scientific degrees at the Technical University - Sofia.

The doctoral student has carried out well-organized research work, has reached correct conclusions and, based on this, has made developments of a scientific and applied nature.

Based on this, I recommend to the esteemed scientific jury to award M.Sc. Eng. Georgi Yanev Kotov the educational and scientific degree "Doctor" in the professional field 5.3 "Communication and Computer Engineering" and the scientific specialty "Automated Information Processing and Management Systems".

02.09.2026 r.

**Reviewer:**

***/Prof. R. Trifonov/***