Farrakhov O. V. — PhD in Engineering, Leading Researcher, Center for Information-analytical and Technical Support of Nuclear Power Facilities Monitoring of the National Academy of Sciences of Ukraine, 34A, Academician Palladin Ave, Kyiv, Ukraine, 03142; farrakhov@ukr.net; ORCID: 0000-0003-4988-126X

Martseva L. A. — D. Sc. in Pedagogy, assistant professor, Zhytomyr Polytechnic State University, 103, Chudnivsyka Str., Zhytomyr, Ukraine, 10005; l.a.martseva@gmail.com; ORCID: 0000-0001-5037-6565

Kotsiubynskyi A. O. — PhD in Physics and Mathematics, Assistant Professor, Ivano-Frankivsk National Technical University of Oil and Gas, 15, Karpatska Str., Ivano-Frankivsk, Ukraine, 76019; Radijrlife@gmail.com; ORCID: 0000-0003-1135-3568

Vlasenko O. V. — Senior Lecturer at the Department of Software Engineering, Zhytomyr Polytechnic State University, 103, Chudnivsyka Str., Zhytomyr, Ukraine, 10005; oleh.vls@gmail.com; ORCID: 0000-0001-6697-2150

http://doi.org/10.35668/2520-6524-2024-3-13
УДК 004.054:[004.032.26+004.85]

Я. О. ІСАЄНКОВ, аспірант
О. Б. МОКІН, д-р техн. наук, проф.

МЕТОД ОЦІНКИ ЧАСТКОВО ЗГЕНЕРОВАНИХ ДАНИХ

Резюме. *Останніми роками генеративні моделі, зокрема автокодувальники, генеративні змагальні мережі та дифузійні моделі, стали невіддільною частиною інновацій у різних галузях, таких, як мистецтво, дизайн, медицина тощо. Завдяки здатності створювати нові зразки даних, вони відкривають широкі можливості для автоматизації та вдосконалення процесів. Однак оцінка якості згенерованих даних залишається складним завданням, оскільки традиційні методи не завжди адекватно відображають різноманітність і реалістичність створених зразків. Зокрема це стосується часткового генерування даних, де зміни застосовуються лише до окремих частин зображення, що значно ускладнює оцінку їх якості.*

У цій статті розглянуто різні підходи до оцінки генеративних моделей, зокрема такі автоматичні метрики, як Inception Score і Fréchet Inception Distance, влучність, повнота, щільність і покриття, а також метод із залученням людини HYPE. Хоча ці метрики добре зарекомендували себе в оцінюванні результатів традиційного генерування, їх використання у випадку частково згенерованих даних може бути недоцільним через їх обмеження. Для розв'язання цієї проблеми в статті запропоновано новий метод оцінювання частково згенерованих даних із залученням людини. Цей метод базується на аналізі трансформованих зображень користувачами, які визначають зони, що зазнали змін, і оцінює їхню якість за допомогою метрик влучності, повноти, F1-міри, шукаючи перетини між реальними зонами та вибраними користувачем із використанням IoU. Запропонований підхід забезпечує більш об'єктивну оцінку реалістичності та якості згенерованих фрагментів зображень під час трансформацій.

Наведено практичний приклад застосування розробленого методу на наборі даних панорамних стоматологічних знімків, де оцінювалася якість трьох моделей: 1) ГЗМ на основі U-генератора; 2) та сама модель, але з післяобробкою вихідного зображення і сегментаційної маски; 3) самовалідована ГЗМ. Оцінку проводили 30 осіб. Середні значення F1-міри для цих моделей становили 0,78, 0,27 і 0,20 відповідно. Оскільки нижчі значення F1-міри в цьому випадку свідчать про кращі результати (чим точніше користувачі ідентифікували трансформації, тим гірше працювала модель), найкращою моделлю за цією метрикою є самовалідована ГЗМ, що також підтверджується суб'єктивними оцінками, зазначеними в працях авторів.

Ключові слова: *аугментація, генерування даних, генеративна змагальна мережа, ГЗМ, комп'ютерний зір, глибоке навчання, самовалідована ГЗМ, оцінювання, нейронні мережі.*

ВСТУП

Генеративні мережі, зокрема автокодувальники [1], генеративні змагальні мережі (ГЗМ) [2], дифузійні моделі [3] стали потужним інструментом для створення таких нових зразків даних, як зображення, текст або навіть музика. Їх широке застосування охоплює різні галузі, включаючи мистецтво, дизайн, медицину.

Оцінка якості згенерованих даних є критично важливою для забезпечення надійності та відповідності згенерованих результатів очікуванням.

Оцінка генеративних мереж є непростим завданням, оскільки згенеровані дані можуть бути дуже різноманітними, а їхня якість не завжди очевидна. З одного боку, важливо враховувати здатність моделі генерувати реалістичні та унікальні зразки, які не є просто копіями тренувальних даних. З іншого боку, необхідно забезпечити, щоб згенеровані дані відповідали певним критеріям, з-поміж яких консистентність стилю, різноманітність і коректність.

Існує декілька підходів до оцінки генеративних мереж, кожен з яких має свої переваги та обмеження. Вибір методів оцінки залежить від конкретного завдання, типу згенерованих даних і вимог до них. Методи оцінки можна розділити на дві категорії: автоматичні та із залученням людини.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Наявні автоматичні методи оцінювання якості згенерованих даних. Однією з найвідоміших метрик є Inception Score (IS) [4]. Її реалізація базується на використанні однойменної нейронної мережі InceptionNet [5], навченої на наборі даних ImageNet [6]. Метрика заснована на двох критеріях: вірність (fidelity), що відповідає за якість створених даних; різноманітність (diversity), що відповідає за те, наскільки різні дані може створювати модель.

Щоб оцінити згенеровані дані (у цьому випадку — лише зображення) найперше вони подаються на вхід до нейронної мережі InceptionNet. На виході отримуються ймовірності приналежності даних до одного з тисячі класів, що представлені в ImageNet (**рис. 1**).

Щоб мати гарні показники вірності конкретні приклади згенерованих даних мають належати лише до одного вихідного класу (класифікаційна модель повинна чітко розуміти, що зображено на картинці серед того, що вона знає), тобто мати низьку ентропію. Для критерію різноманітності потрібно, щоб набір згенерованих даних охоплював якомога більшу кількість класів, тобто мав високу ентропію за класами. Комбінація цих метрик утворює фінальну оцінку: чим більшою буде різниця в цих розподілах, тим вищим буде значення IS, що розраховується за допомогою міри розходження Кульбака — Лейблера (**рис. 2**).

Одним із недоліків IS є невикористання інформації про реальні дані та їх розподіли. На цій ідеї побудована метрика Fréchet Inception Distance (FID) [7]. FID показує те, наскільки розподіли реальних і згенерованих даних близькі між собою за відстанню Фреше.

Щоб порівнювати розподіли кожне зображення мусить бути представлено у вигляді вектора. Оскільки немає сенсу порівнювати між собою пікселі картинок, то в цьому підході, як і у IS, використовується нейронна мережа InceptionNet для отримання абстрактних векторів ознак з останніх шарів моделі, оскільки ці шари описують важливі з точки зору моделі ознаки вхідних зображень різних об'єктів. Для таких реальних і згенерованих векторів даних

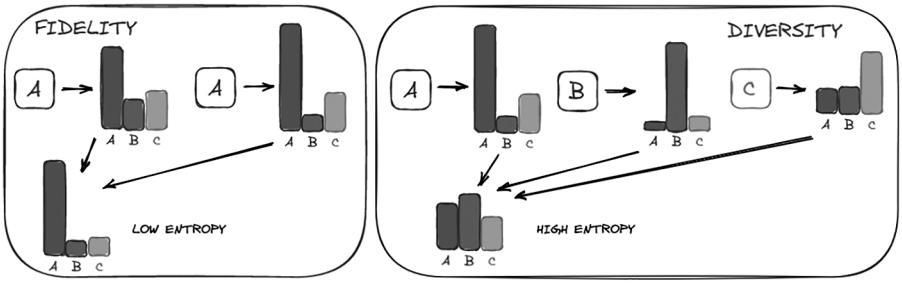

Рис. 1. Принцип роботи Inception Score

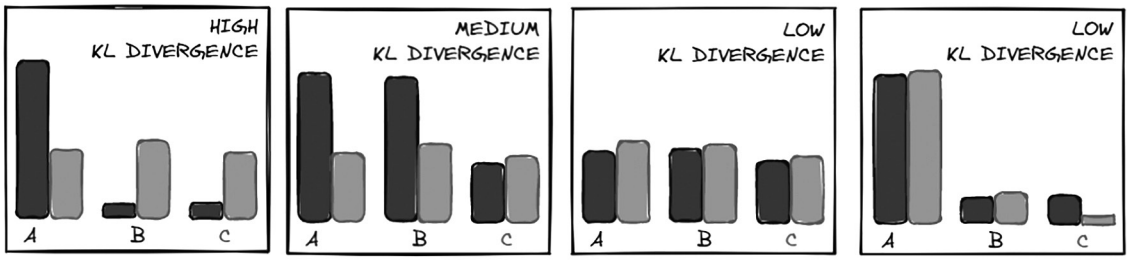

Рис. 2. Оцінка близькості розподілів

розраховуються значення середнього та коваріації. Далі отримані розподіли порівнюються між собою через відстань Фреше, що схематично показано на **рисунку 3**.

Влучність (precision) та повнота (recall) зазвичай застосовуються як вагомі метрики в задачах класифікації для підрахунку частки позитивних зразків серед знайдених і частки загального числа знайдених позитивних зразків відповідно. Влучність та повнота також можуть застосовуватися під час оцінювання згенерованих даних.

На **рисунку 4b** показано, що влучність — це ймовірність того, що згенеровані дані будуть у межах розподілу реальних даних, що є визначенням вірності. Якщо більшість згенерованих даних попадають у область, що займають реальні дані, то це може свідчити про гарну якість зображень. Відповідно, якщо згенеровані дані знаходяться поза межами реальної області, то це може свідчити про погану якість цих даних.

На **рисунку 4c** показана оцінка повноти, що відповідає ймовірності того, що реальні дані будуть входити в згенерований розподіл, за що відповідає різноманітність. Якщо всі реальні дані знаходяться в межах згенерованого розподілу, то це є свідченням того, що згенеровані дані мають дуже різні варіації. Якщо згенеровані охоплюють лише маленьку частину реальних даних, то це свідчить про те, що такі дані практично не мають різноманітності та всі дуже схожі між собою, що також називається “колапс режиму” (mode collapse) (**рис. 5**).

Щільність (density) та покриття (coverage) — це метрики, що є аналогами влучності та повноти, але вони розв’язують ряд проблем, притаманних наведеним вище метрикам, а саме: мають можливість ідентифікувати ідентичні реальний і згенерований розподіли, показуючи близькі до 100 % значення; є стійкими до аномалій — як у згенерованих даних, так і в реальних; мають чіткий і простий алгоритм вибору гіперпараметрів оцінки [10].

Порівняння щільності та покриття з влучністю та повнотою показано на **рисунку 6**. З **рисунка 6а** видно, що, якщо в реальних даних є аномальні значення, то вони можуть дуже розширити покриття простору реальних даних та охопити погано згенеровані дані, причому піднявши влучність до 100 %. Щільність у такій ситуації є кращим показником, оскільки ця метрика враховує в скільки зон потрапляють згенеро-

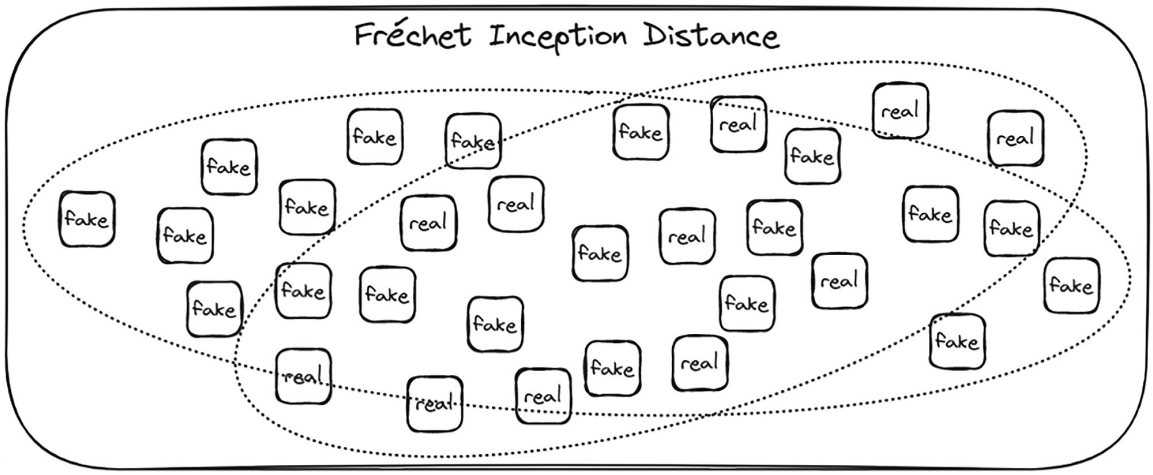

Рис. 3. Схематичне порівняння розподілів даних

Рис. 4. Влучність та повнота [8]. P_g — згенеровані дані, P_r — реальні дані.

вані дані відносно кількості реальних даних у цій зоні (кількість реальних даних у кожній області є одним із гіперпараметрів алгоритму, зони збільшуються, допоки кількість реальних даних не досягне заданого значення; на **рисунку 6a** це значення дорівнює двом). У прикладі чотири з п'яти фейкових семплів покривається лише однієї реальною областю, що є аномальною.

Ліва частина **рисунка 6b** показує, що якщо генеративна модель має погану якість, але гарну різноманітність, то згенеровані дані можуть охоплювати великий простір і піднімати значення метрики повноти до 100 %. Водночас метрика покриття створює області на основі справжніх прикладів і підраховує, який відсоток справжніх прикладів має щонайменше один несправжній приклад у своїй області. Із рисунка видно, що оцінка згенерованих даних вийшла коректною та низькою.

Метрики GAN-train та GAN-test [11] використовуються для вимірювання різноманітності та реалізму даних відповідно. У першому підході створюється умовна генеративна мережа (ГМ) на основі тренувальних даних і завдяки цій моделі створюється новий набір синтетичних даних (з відомими для них класами). На цих даних тренується класифікатор. Останнім кроком є оцінка отриманого класифікатора на окремих реальних тестових даних. Якщо класифікатор має гарну точність за всіма категоріями, то метрика різноманітності буде висока, адже генератору вдалося створити такі дані, що інша, натренована на цих даних модель, змогла чітко розрізняти реальні дані. Якщо згенеровані зразки будуть погано описувати реальні дані, то і класифікаційна модель не зможе з гарною точністю розрізняти тестовий набір. Схематично цей процес показано на **рисунку 7**.

Підхід GAN-test складається зі схожих компонентів, але відрізняється процесами та зв'язками між ними. Спочатку створюється умовна ГМ та класифікатор на наявних зразках тренувального набору даних. Далі генератором створюється набір синтетичних даних. Завдяки класифікатору ці синтетичні дані оцінюються: якщо мітка класифікатора з високою ймовірністю збігається з міткою, яку було передано моделі-генератору, то ці дані вважаються високоякісними, оскільки навчений на реальних даних класифікатор сприймає їх. Якщо ж модель-генератор буде створювати дані поганої якості, то, найбільш імовірно, класифікатор не

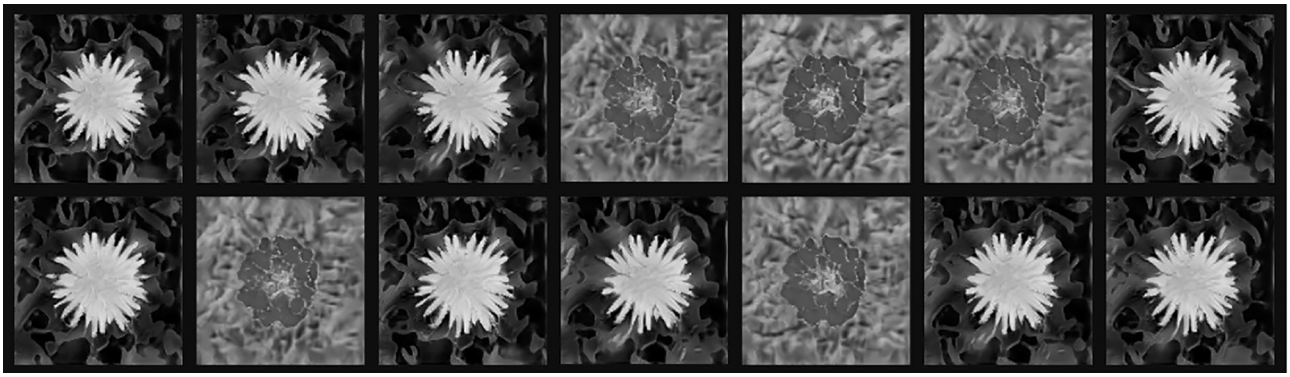

Рис. 5. Приклад явища колапс режиму [9]

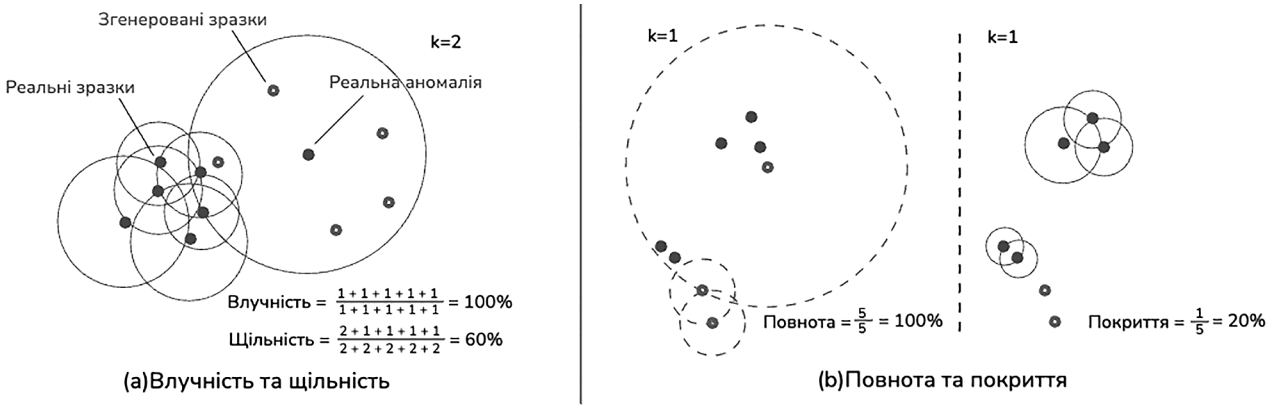

Рис. 6. Щільність та покриття [10]

зможе розрізнити в них потрібні йому ознаки. Схематично цей процес показано на **рисунку 8**.

Ці два підходи зазвичай використовуються разом, тобто кроки створення ГМ і генерування даних робляться лише раз, а потім створюються вже класифікатори та здійснюється оцінювання за конкретною метрикою.

Ще однією проблемою, якій варто приділити увагу, є копіювання даних. Генеративна модель може просто запам'ятати (частково або навіть повністю) частини тренувальних даних і відтворювати їх. Непараметричний тест для виявлення копіювання даних [12] — це метод, що дає змогу визначати чи не перенавчилася генеративна модель на тренувальному наборі. Для цього методу потрібні тренувальний, тестовий і згенерований набори даних. Якщо згенеровані дані, наприклад, у просторі ознак моделі InceptionNet, в середньому ближче до тренувального набору, ніж до тестового, це може свідчити про проблему копіювання даних. Також варіація цього методу оцінки передбачає ситуації, коли копіювання вийшло лише на частині тренувальних даних. Для цього простір ділиться на області і на кожній з областей проводиться оцінювання ступеня копіювання.

Perceptual Path Length (PPL) [13] допомагає оцінити те, наскільки складним є латентний простір після навчання генеративної моделі. Якщо зробити інтерполяцію в цьому просторі, то очікується, що в проміжних зображеннях буде схожий набір ознак, що наявний і у крайніх зображеннях, у ситуації, якщо простір є не викривленим. Але якщо простір є занадто заплутаним і складним, то в процесі інтерполяції можуть з'являтися дуже незвичайні ознаки на проміжних зображеннях. PPL дозволяє кількісно оцінити те, наскільки змінюються ознаки даних у процесі інтерполювання в просторі.

Метрика лінійної роздільності [13], як і PPL, оцінює якість латентного простору. Якщо простір є не дуже складним в ньому може з'явитись вектори, що відповідають за присутність тієї чи іншої ознаки. Ця метрика надає можливість оцінити таку здібність простору, надаючи оцінку стосовно того, наскільки добре точки прихованого простору можуть бути розділені на два кластери за допомогою лінійної гіперплощини,

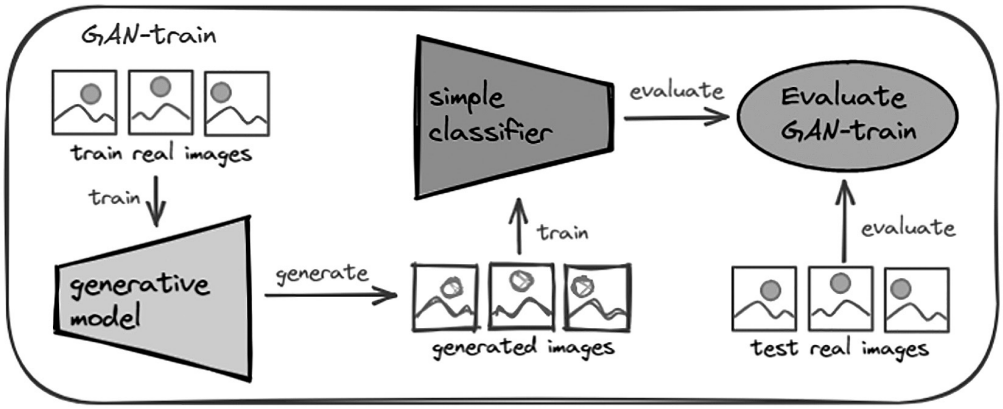

Рис. 7. Схема підходу GAN-train

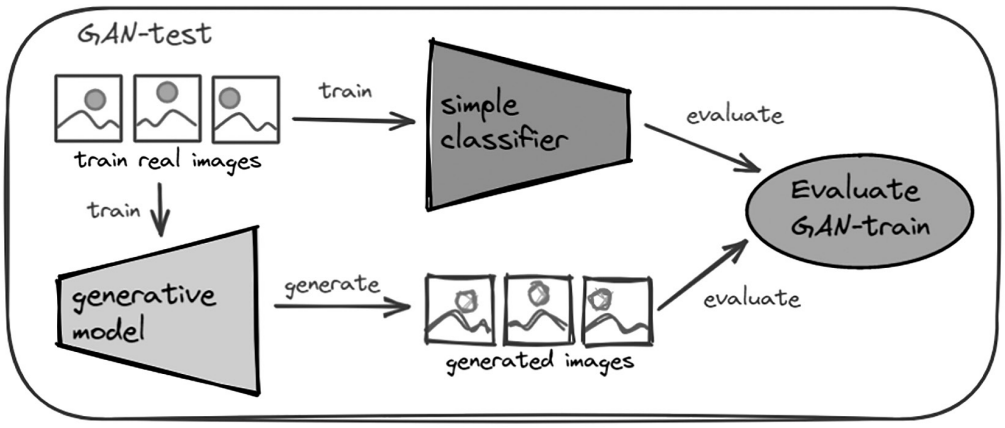

Рис. 8. Схема підходу GAN-test

так щоб кожен кластер відповідав певному бінарному атрибуту даних.

Наявні методи оцінювання згенерованих даних із залученням людини. Методи, які передбачають залучення до оцінювання людини, мають перевагу в тому, що людина краще може розуміти деякі нюанси предметної області та світу (наприклад, дуже легко впізнають неправильні риси обличчя, несумісності в згенерованому ландшафті тощо).

Часто для оцінювання моделей потрібно згенерувати деякий набір даних і віддати ці дані на оцінку групі експертів, які ретельно переглянуть дані та дійдуть висновку про те, наскільки доброю є якість даних. Однак цей процес повинен мати чіткі специфікації та критерії якості.

Особливості залучення людини до оцінювання згенерованих даних найкраще описані в методі-стандарті Human eYe Perceptual Evaluation (HYPE) [14]. В основі цього стандарту є вимірювання реалістичності сприйняття результатів людиною, чітка послідовність дій при оцінюванні, забезпечення стабільності отримання результатів, а краудсорсингова платформа дає змогу ефективно використовувати грошові та часові ресурси.

Існує дві варіації стандарту HYPE: часовий ($HYPE_{time}$) та необмежений у часі ($HYPE_{\infty}$). Часова варіація полягає в тому, що вираховується час, який потрібен людині, щоб зрозуміти справжність або несправжність зображення. Наприклад, оцінка в одну секунду показує, що людині потрібно не менше однієї секунди, щоб відрізнити реальне зображення від згенерованого.

Процедура оцінки базується на адаптивному методі сходинок: зображення впродовж деякого часу показується людині; потім людині потрібно відповісти на запитання про справжність зображення; якщо відповідь правильна, то наступного разу часу буде трохи менше, а якщо — неправильна, то часу буде більше. Зазвичай використовуються часові надбавки у відношенні три до одного: за правильну відповідь віднімається в три рази більше часу, аніж додається в разі неправильної відповіді. Для кінцевої оцінки часу зазвичай обирається 150 зображень, половина з яких реальні, а половина — згенеровані, а межі часової оцінки становлять від 100 мс до 1 с.

На основі $HYPE_{time}$ побудовано метод $HYPE_{\infty}$, який є більш простим, швидким і дешевим. Цей стандарт відійшов від часових рамок і дозволяє людині вирішувати реальне чи згенероване зображення їй показано без прив'язки до часу. Зазвичай використовується 50 реальних і 50 згенерованих зображень. 10 % $HYPE_{\infty}$ показує, що лише 10 % зображень є незрозумілими для експертів, 50 % — показує, що людина плутає половину зображень, тобто згенеровані зображення не можна відрізнити від реальних, а більше 50 % — показує, що згенеровані зображення є гіперреалістичними — схожі на реальні більш ніж справжні. У такому разі людина дає мітку справжніх зображень згенерованим.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Використання FID для частково згенерованих даних. Часткове генерування даних [15] має декілька значних переваг над звичайним генеруванням даних. Однією з переваг є можливість ефективно трансформувати зображення високої роздільної здатності, не вимагаючи значних обчислювальних ресурсів. Це досягається завдяки тому, що зміни застосовуються лише до окремих частин зображення, а не до всього зображення цілком. Розглянемо приклад зображення з роздільною здатністю 1920×1080 пікселів. Замість того, щоб трансформувати повністю все зображення, аугментація може бути застосована лише до невеликої його частини, наприклад, до області розміром 256×256 пікселів. Це дає змогу значно зменшити обчислювальне навантаження та прискорити процес підготовки даних.

Також часткова аугментація дозволяє зберегти основну структуру та важливі характеристики вихідного зображення, водночас збільшуючи різноманітність набору даних. Це особливо корисно у випадках, коли важливо зберегти контекстну інформацію та дозволяє більш гнучко підходити до процесу підготовки даних. Наприклад, можна вибірково трансформувати лише ті області зображень, які потребують додаткової варіативності для покращення навчання моделі. Уявімо собі задачу розпізнавання облич на зображеннях високої роздільної здатності. Замість того, щоб аугментувати все зображення, можна застосувати трансформації лише до облич або окремих частин обличчя (очей, носу, роту), що значно скоротить час обробки та збереже важливу контекстну інформацію. Наприклад, обертання, додавання шуму чи зміна яскравості може бути застосовано лише до невеликої області навколо очей, тоді як решта зображення залишиться незмінною.

Разом із перевагами часткова аугментація даних додає викликів. Наприклад, під час оцінювання якості трансформацій зображення за допомогою такого методу, як FID, можуть виникати труднощі з виявленням значних відмінностей, якщо аугментований фрагмент є невеликим. Це добре демонструє наведений нижче приклад. На **рисунку 9** зображено чотири рентгенівські знімки. Верхнє ліве зображення є оригінальним, тоді як на решті зображень

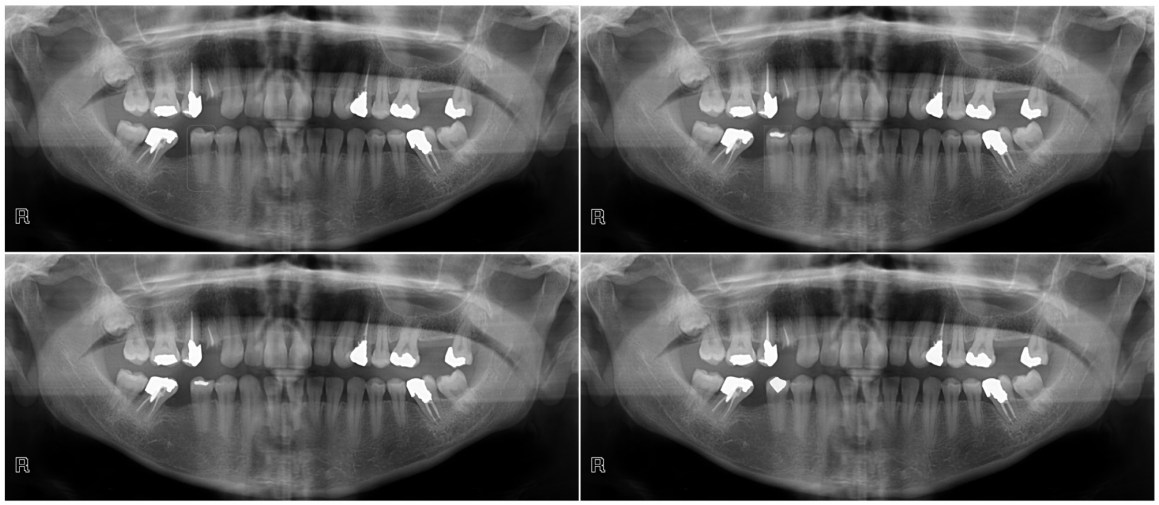

Рис. 9. Приклад часткового генерування з використанням різних підходів

трансформовано область, що виділена на першому зображенні червоною рамкою. Кожне з інших зображень є результатом різних моделей трансформацій.

Зображення у верхньому правому куті є результатом генеративної моделі на основі U-GAN [15], яка сильно спотворює вхідний фрагмент зображення, роблячи його вигляд ненатуральним і не даючи змогу правильно додати його в контекст знімка. Два зображення внизу отримані за допомогою більш якісних методів: ліве — за допомогою агрегації вхідного зображення та маски від першої моделі [15], а праве — з використанням самовалідованої генеративної змагальної мережі [16].

Очікується, що метрика якості має відображати кількісно суттєві дефекти в поганої моделі. Для перевірки цього було взято 20 зображень і для кожного з них трансформовано 1–2 зуби. Додатково взято 10 зображень без трансформації, щоб оцінити те, наскільки сильно відрізняються метрики та розподіли в порівнянні з узагалі іншими зображеннями.

На **рисунку 10** показано попарне значення FID оцінки для кожного набору зображень:

- input images — нетрансформовані зображення;
- model_1 — трансформовані вхідні зображення з використанням першої моделі на основі U-GAN;
- model_2 — трансформовані вхідні зображення з використанням другої моделі з використанням післяобробки;
- model_3 — трансформовані вхідні зображення з використанням третьої моделі самовалідованої ГЗМ;
- additional images — додатковий набір нетрансформованих зображень.

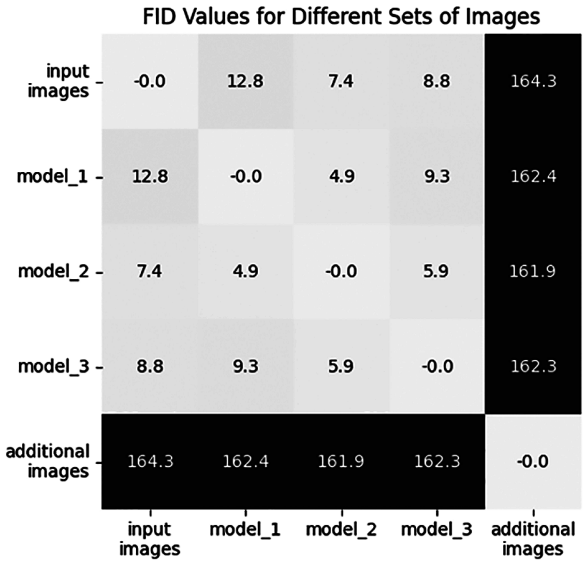

Рис. 10. Попарне значення FID для різних наборів даних

Як показано на зображенні, немає чіткої різниці між добре аугментованими зображеннями, погано аугментованими та оригінальним зображенням. Проте додаткові зображення кардинально відрізняються від усіх наборів даних.

На **рисунку 11** показано, який розподіл мають звичайні зображення та додаткові звичайні у двовимірному просторі, отриманому за допомогою методу головних компонент (PCA). Зображення були оброблені моделлю InceptionV3 для отримання векторів ознак, що мають розмірність 2048. Після цього було застосовано PCA, що дав змогу знизити розмірність векторів ознак із 2048 до двох основних компонент.

На лівій частині **рисунка 11** показано розподіл реальних даних у двовимірному просторі

(кожна біла точка — окреме зображення). На правій панелі зображено розподіл додаткових зображень. Ці два розподіли мають досить різні параметри і форму відповідно.

Аналогічні розподіли показано на **рисунку 12**, але кожен із розподілів є результатом різної моделі генерування. Усі три розподіли мають дуже близьку природу до розподілу вхідних даних **рисунку 11**.

З наведеного вище стає очевидним, що FID та інші методи екстрагування вектора ознак із цілого зображення для оцінювання частково згенерованих даних не підходять для оцінювання часткового згенерованих даних.

Метод оцінки частково згенерованих даних. У статті пропонується новий метод оцінювання частково згенерованих даних із залученням людини. На відміну від методу HYPE, цей метод спрямований на акцентування уваги саме на згенерованих частинах зображення і на те, наскільки добре людина здатна ідентифікувати ці частини.

Сутність методу полягає в тому, що людина аналізує трансформоване зображення та ідентифікує змінені зони. Зміни можуть бути численними і різноманітними, тому завданням є знайти і визначити всі з них. Головні принципи такі:

- якщо людина правильно ідентифікує всі змінені ділянки, то вона отримує максимальну оцінку, а модель, яка згенерувала зміни, — мінімальну.
- якщо людина не визначає жодної зміненої ділянки, то вона отримує мінімальну оцінку, а модель — максимальну.

Для об'єктивного оцінювання результатів використовуються декілька ключових метрик.

1. Влучність (Precision). Ця метрика показує, яка частка змін, знайдених людиною, дійсно є зміненими. Висока влучність свідчить про те, що експерт в основному правильно ідентифікує зміни з мінімальною кількістю хибних висновків

$$Precision = \frac{TP}{TP+FP}, \qquad (1)$$

де *TP* — кількість правильно знайдених зон, *FP* — кількість знайдених зон, що є помилковими.

2. Повнота (Recall). Ця метрика відображає, яка частка всіх змінених ділянок була знайдена людиною. Висока повнота свідчить про те, що

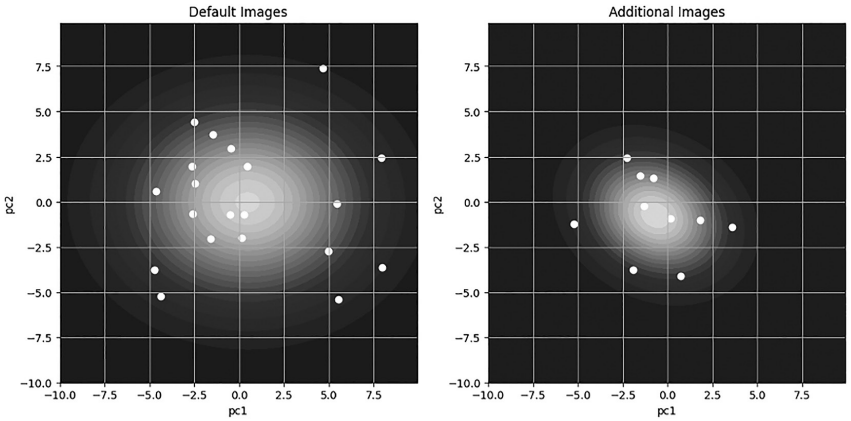

Рис. 11. Розподіли двох наборів реальних даних

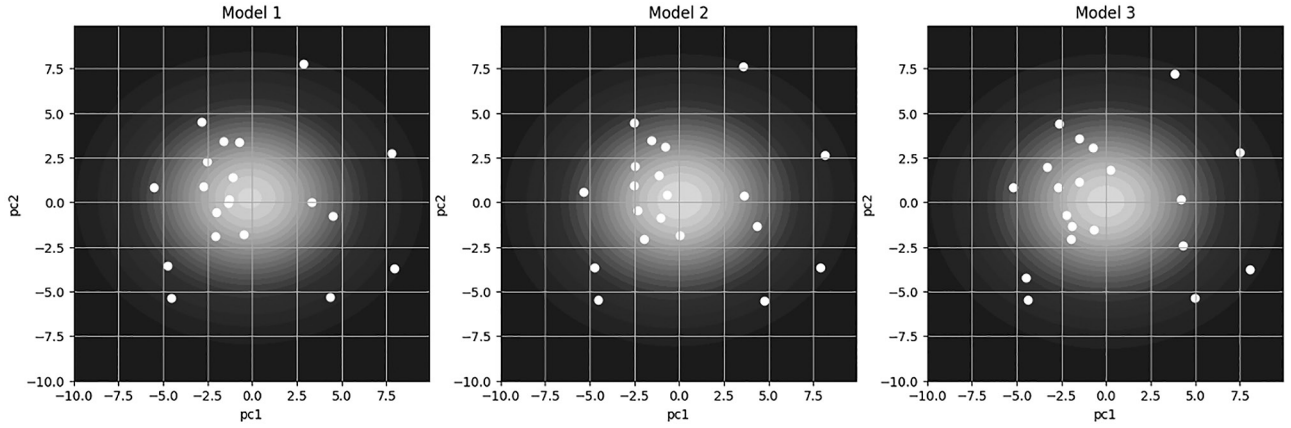

Рис. 12. Розподіли частково згенерованих наборів даних

експерт знаходить більшість або всі зміни, зроблені моделлю:

$$Recall = \frac{TP}{TP+FN}, \quad (2)$$

де *TP* — кількість правильно знайдених зон, *FP* — кількість незнайдених зон.

3. F1-міра (F1 score). Це гармонійне середнє між точністю та повнотою, яке забезпечує баланс між цими двома показниками. Висока F1-міра вказує на те, що і влучність, і повнота знаходяться на високому рівні:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \quad (3)$$

Для оцінки того, наскільки точно користувач визначив змінені ділянки, використовується метрика *IoU* (Intersection over Union). Ця метрика визначає ступінь перетину області, вказаної користувачем, і реально згенерованої області. Чим більше співпадіння між цими областями, тим вищий показник *IoU*, що свідчить про більш точну ідентифікацію змін:

$$IoU = \frac{A \cap B}{A \cup B}, \quad (4)$$

де *A* — справжня зона трансформації, *B* — зона, вибрана користувачем.

Алгоритм роботи розробленого методу, що показано на **рисунку 13**, охоплює наступні етапи:

1) вибір зображень для оцінювального набору даних та вибір зон для трансформації (пунктирна рамка). На першому етапі з вибраних зображень визначаються ділянки, які підлягатимуть трансформації;

2) трансформація вибраних ділянок генеративною моделлю. Кожне зображення піддається обробці генеративною моделлю, яка трансформує вибрані ділянки. У результаті отримуються трансформовані зображення та цільові координати боксів, де відбулася трансформація;

3) оцінка трансформованих зображень користувачами. Трансформовані зображення надаються групі користувачів. Їм необхідно ідентифікувати ділянки, які вони вважають аугментованими (звичайні рамки на зображенні), тобто такими, що зазнали змін;

4) розрахунок метрик для кожного користувача і зображення. На основі точності визначення ділянок, що були аугментовані на конкретному зображенні, для кожного користувача розраховуються індивідуальні метрики;

5) агрегація результатів серед усіх користувачів. Результати оцінки всіх експертів агрегуються для отримання метрики для кожного окремого зображення. Ця метрика відображає загальну оцінку якості трансформації наданої генеративною моделлю для одного зображення;

6) агрегація результатів по всім картинкам. На цьому останньому етапі відбувається агрегація результатів по всьому набору зображень для отримання загальної оцінки. Це значення є ключовим показником ефективності роботи генеративної моделі.

Практичне використання розробленого методу. На **рисунку 14** показаний приклад роботи онлайн-застосунку для перевірки розробленого підходу. Користувачі мають можливість виділяти фрагменти на рентгенівському знімку зубів, які вони вважають трансформованими чи модифікованими. Ці виділені області вибираються прямокутниками на знімку. Після завершення роботи з поточним зображенням користувач може перейти до наступного, натиснувши на кнопку “Next”, що розташована внизу інтерфейсу.

В експерименті взяли участь 30 користувачів, кожному з яких необхідно було опрацювати по 20 зображень трансформованих кожною

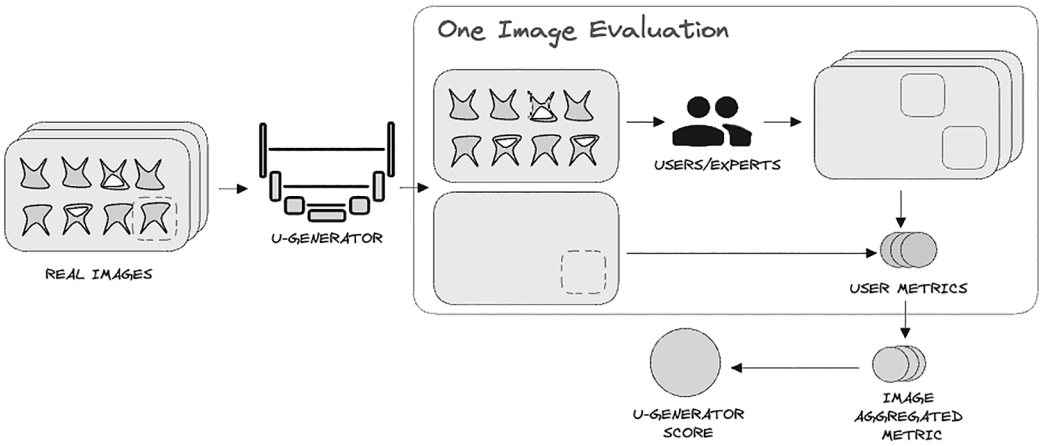

Рис. 13. Схема методу оцінки частково згенерованих даних

з трьох зазначених вище моделей. Додатково додано 10 зображень, які не були трансформовані зовсім. На кожному зображенні можуть бути трансформовані від 0 до 2 зубів.

На **рисунку 15** представлено гістограми, які відображають розподіли значень F1-метрики для трьох моделей, що тестувалися на 30-ти користувачах. Можна зазначити, що користувачі здобували найкращі результати на моделі 1, що є найслабшою. Розподіли метрики для моделі 2 та моделі 3 набагато кращі, розподіл для моделі 3 зміщений трохи лівіше, що говорить про перевагу над моделлю 2. Нагадаємо, що нижче значення F1 вказує на кращу якість моделі, тобто користувачу складніше відрізнити згенеровані частини від реальних.

Далі наведено середні значення метрик точності, повноти та F1-метрики для всіх користувачів по кожній із моделей:

1) модель 1 показала найгірші середні значення за всіма метриками: влучність — 0,785556, повнота — 0,783333, F1 — 0,776167;

2) модель 2 мала помітно кращі показники, середні значення яких були такі: влучність — 0,304167, повнота — 0,267500, F1 — 0,274500;

3) модель 3 виявилася найкращою серед усіх трьох моделей із середніми значеннями: влучність — 0,220000, повнота — 0,193333, F1 — 0,199444.

На **рисунках 16–18** показані всі вибори користувачів для першої та третьої моделі. Центри трансформованих боксів зубів показані знаком “X”.

На **рисунку 16** зліва видно, що неякісна трансформація зуба верхньої щелепи сильно привернула увагу користувачів і він вибирався дуже багато разів. З правої сторони видно, що цей зуб, хоч був знайдений низкою користувачів, але вони вибирали і не трансформовані зуби на цій картинці.

Ліва сторона **рисунка 17** показує, що два підозрілих зуби у верхній щелепі мали сильні ознаки трансформації, тому інші дивні пломби

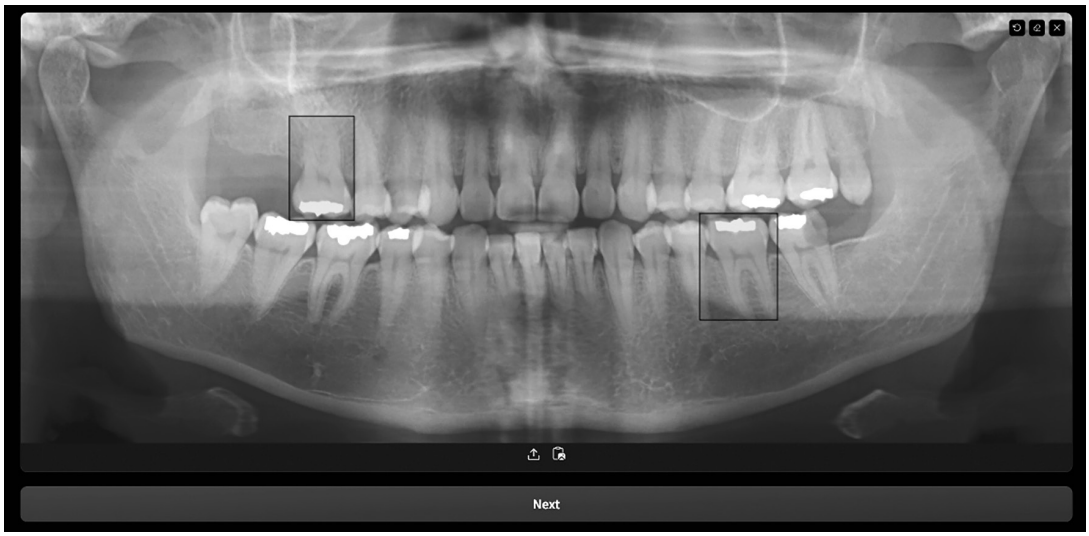

Рис. 14. Екранна форма застосунку розробленого методу

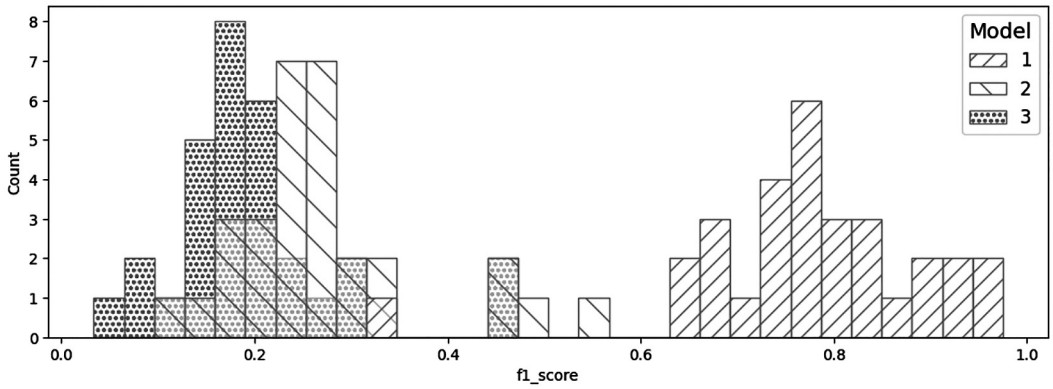

Рис. 15. Розподіли F1 оцінок користувачів

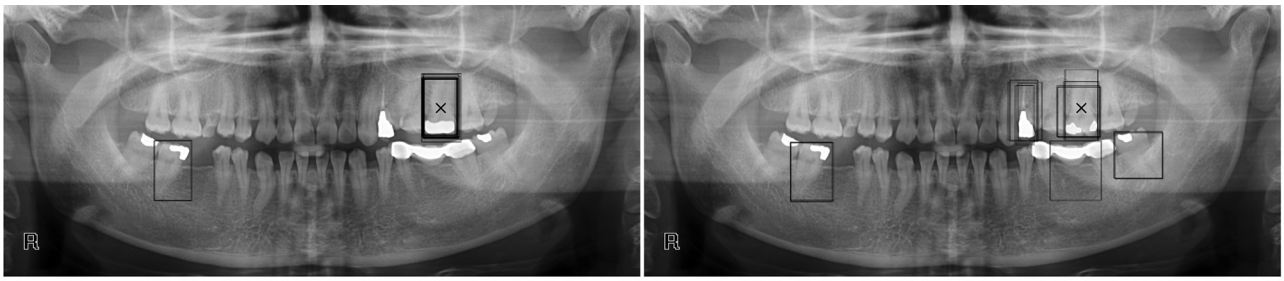

Рис. 16. Трансформація 27-го зуба моделлю 1 та 3

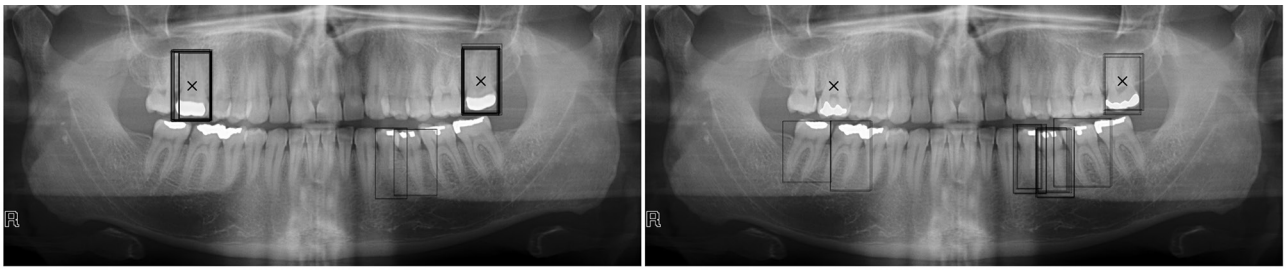

Рис. 17. Трансформація 16-го та 27-го зубів моделлю 1 та 3

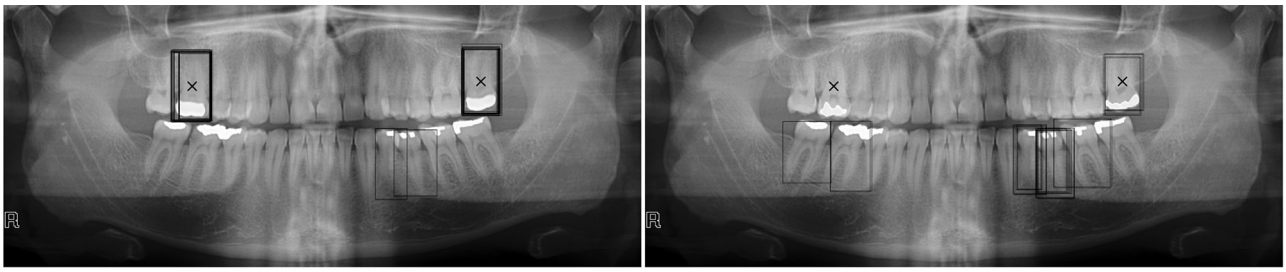

Рис. 18. Трансформація 26-го зуба моделлю 1 та 3

нижньої щелепи не вибиралися часто. Модель 3 не створила ознак трансформації, завдяки чому один із трансформованих зубів взагалі не був обраний, а також було обрано багато разів дуже підозрілі пломби з нижньої щелепи.

Рисунок 18 демонструє дуже помітний артефакт у апріорі поганої моделі, за рахунок якого більшість користувачів сконцентрували увагу на ньому. Натомість самовалідована ГЗМ створила гарний екземпляр і користувачі розсіяли свою увагу на різні нетрансформовані екземпляри.

ВИСНОВКИ

У пропонованій увазі статті розглянуто проблему оцінювання згенерованих даних. Проаналізовано наявні методи та підходи (зокрема автоматичні та із залученням людини) до оцінювання результатів традиційного генерування, коли зображення створюється повністю і потрібно визначити чи воно справжнє, чи згенероване. На прикладі Fréchet Inception Distance показано неефективність стандартних підходів для оцінки частково згенерованих зображень, тобто таких, у яких згенерованою є лише певна частина.

У статті запропоновано новий метод оцінювання частково згенерованих даних із залученням людини, де користувачам потрібно знаходити на зображенні трансформовані частини. Для оцінки вибору користувачів використана F1-міра. Чим кращу оцінку мають користувачі, тим гіршу отримає модель, що свідчить про те, що модель трансформує дані занадто погано, і користувачі без проблем знаходять ці частини.

Наведено практичний приклад застосування розробленого методу на наборі даних панорамних стоматологічних знімків, який показав ефективність розроблено підходу. У тестуванні взяли участь 30 людей, які оцінювали якість трьох моделей: 1) ГЗМ на основі U-генератора; 2) та сама модель, але з післяобробкою вихідного зображення; 3) самовалідована ГЗМ. У результаті самовалідована ГЗМ отримала суттєво кращі результати.

Використання розробленого методу оцінювання може допомогти у виборі кращих моделей у задачі часткової аугментації даних для задач глибокого навчання.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Kramer M*. Nonlinear principal component analysis using autoassociative neural networks / M. Kramer // AIChE. — 1991. — Vol. 37. — No. 2. — P. 233–243. DOI: https://doi.org/10.1002/aic.690370209.
2. *Goodfellow I.* Generative adversarial networks / I. Goodfellow, J. Pouget-Abadie, M. Mirza et al. // Communications of the ACM. — 2020. — Vol. 63. — No. 11. — P. 139–144. DOI: https://doi.org/10.1145/3422622.
3. *Chen M.* An overview of diffusion models: Applications, guided generation, statistical rates and optimization / M. Chen, S. Mei, J. Fan, M. Wang // arXiv e-prints. — 2024. DOI: https://doi.org/10.48550/arXiv.2404.07771.
4. *Salimans T.* Improved Techniques for Training GANs / T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung et al. // NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems. — 2016. — P. 2234–2242. DOI: https://doi.org/10.48550/arXiv.1606.03498.
5. *Szegedy C.* Going deeper with convolutions / C. Szegedy, Wei Liu, Yangqing Jia et al. // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — Boston, MA, USA, 2015. — P. 1–9. DOI: https://doi.org/10.1109/CVPR.2015.7298594.
6. *Russakovsky O*. ImageNet Large Scale Visual Recognition Challenge / O. Russakovsky, J. Deng, H. Su, et al. // International Journal of Computer Vision. — 2015. — Vol. 115. — P. 211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y.
7. *Heusel M.* GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium / M. Heusel, H. Ramsauer, T. Unterthiner et al. // NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. — 2017. — P. 6629–6640. DOI: https://doi.org/10.48550/arXiv.1706.08500.
8. *Sajjadi M. S.* Assessing generative models via precision and recall / M. S. Sajjadi, O. Bachem, M. Lucic et al. // Advances in Neural Information Processing Systems. – 2018. — P. 5228–5237.
9. *Ісаєнков Я. О.* Аналіз генеративних моделей глибокого навчання та особливостей їх реалізації на прикладі WGAN / Я. О. Ісаєнков, О. Б. Мокін // Вісник Вінницького політехнічного інституту. — 2022. — № 1. — С. 82–94. — DOI: https://doi.org/10.31649/1997-9266-2022-160-1-82-94.
10. *Naeem M. F.* Reliable fidelity and diversity metrics for generative models / M. F. Naeem, S. J. Oh, Y. Uh et al. // International Conference on Machine Learning. — November 2020. — P. 7176–7185.
11. *Shmelkov K.* How Good Is My GAN? / K. Shmelkov, C. Schmid, K. Alahari // Computer Vision — ECCV 2018. Lecture Notes in Computer Science. — Springer, Cham. — October 2018. — Vol. 11206. — P. 218–234. DOI: https://doi.org/10.1007/978-3-030-01216-8_14.
12. *Meehan C*. A non-parametric test to detect data-copying in generative models / C. Meehan, K. Chaudhuri, and S. Dasgupta // International Conference on Artificial Intelligence and Statistics. — 2020, April.
13. *Karras T*. A Style-Based Generator Architecture for Generative Adversarial Networks / T. Karras, S. Laine, T. Aila // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — Long Beach, CA, USA, 2019. — P. 4396–4405. DOI: https://doi.org/10.1109/CVPR.2019.00453.
14. *Zhou S.* Hype: A benchmark for human eye perceptual evaluation of generative models / S. Zhou, M. Gordon, R. Krishna et al. // Advances in neural information processing systems. — 2019. — No. 32.
15. *Ісаєнков Я. О.* Трансформація цільового класу для задачі сегментації з використанням U-GAN / Я. О. Ісаєнков, О. Б. Мокін // Вісник Вінницького політехнічного інституту. — 2024. –№ 1. — С. 81–87. DOI: https://doi.org/10.31649/1997-9266-2024-172-1-81-87.
16. *Ісаєнков Я. О.* Самовалідований U-GAN для трансформації цільового класу в задачах сегментації / Я. О. Ісаєнков, О. Б. Мокін // Вісник Вінницького політехнічного інституту. — 2024. — № 3. — С. 102–111. DOI: https://doi.org/10.31649/1997-9266-2024-174-3-102-111.

REFERENCES

1. Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative Neural Networks. *AIChE Journal*. *37* (2), 233–243. DOI: https://doi.org/10.1002/aic.690370209.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. et al (2020). Generative Adversarial Networks. *Communications of the ACM*. *63* (11), 139–144. DOI: https://doi.org/10.1145/3422622.
3. Chen, M., Mei, S., Fan, J., & Wang, M. (2024). An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv e-prints*. DOI: https://doi.org/10.48550/arXiv.2404.07771.
4. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems.* P. 2234–2242. DOI: https://doi.org/10.48550/arXiv.1606.03498.
5. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., & Anguelov, et al. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: https://doi.org/10.1109/cvpr.2015.7298594.
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Ma, S. et al (2015). Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*. *115*, 211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y.
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2018). GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems.* P. 6629–6640. DOI: https://doi.org/10.48550/arXiv.1706.08500.
8. Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., & Gelly, S. (2018). Assessing generative models via precision and recall. *Advances in Neural Information Processing Systems.* P. 5228–5237.
9. Isaienkov, Ya. O., & Mokin, O. B. (2022). Analiz heneratyvnykh modelei hlybokoho navchannia ta osoblyvostei yikh realizatsii na prykladi WGAN [Analysis of generative deep learning models and features of their implementation on the example of WGAN]. *Visnyk Vinnytskoho politekhnichnoho instytutu* [Bulletin of the Vinnytsia Polytechnic Institute]. *1,* 82–94. DOI: https://doi.org/10.31649/1997-9266-2022-160-1-82-94 [in Ukr.].
10. Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., & Yoo, J. (2020). Reliable fidelity and Diversity Metrics for generative models. *International Conference on Machine Learning.* P. 7176–7185.
11. Shmelkov, K., Schmid, C., & Alahari, K. (2018). How good is my gan? *Lecture Notes in Computer*

Science. Vol. 11206. P. 218–234. DOI: https://doi.org/10.1007/978-3-030-01216-8_14.
12. Meehan, C., Chaudhuri, K., & Dasgupta, S. (2020). A non-parametric test to detect data-copying in generative models. *International Conference on Artificial Intelligence and Statistics.*
13. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative Adversarial Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. P. 4396–4405. DOI: https://doi.org/10.1109/cvpr.2019.00453.
14. Zhou, S., Gordon, M. L., Krishna, R., Narcomey, A., Fei-Fei, L., & Bernstein, M. S. (2019). Hype: A benchmark for human eye perceptual evaluation of Generative Models. *Advances in neural information processing systems. 32.*
15. Isaienkov, Ya. O., & Mokin, O. B. (2024). Transformatsiia tsilovoho klasu dlia zadachi sehmentatsii z vykorystanniam U-GAN [Target class transformation for segmentation task using U-GAN]. *Visnyk Vinnytskoho politekhnichnoho instytutu* [Bulletin of the Vinnytsia Polytechnic Institute]. *172* (1), 81–87. DOI: https://doi.org/10.31649/1997-9266-2024-172-1-81-87 [in Ukr.].
16. Isaienkov, Ya. O., & Mokin, O. B. (2024). Samovalidovanyi U-GAN dlia transformatsii tsilovoho klasu v zadachakh sehmentatsii [Self-validated U-gan for target class transformation in segmentation tasks]. *Visnyk Vinnytskoho politekhnichnoho instytutu* [Bulletin of the Vinnytsia Polytechnic Institute]. *3,* 102–111. DOI: https://doi.org/10.31649/1997-9266-2024-174-3-102-111 [in Ukr.].

Ya. O. ISAIENKOV, Postgraduate Student
O. B. MOKIN, D. Sc. in Engineering, Professor

METHOD FOR EVALUATING PARTIALLY GENERATED DATA

Abstract. *Generative models, such as autoencoders, generative adversarial networks, and diffusion models, have become an integral part of innovation in various fields in recent years, including art, design, medicine, and more. Due to their ability to create new data samples, they open broad opportunities for automation and process improvement. However, assessing the quality of generated data remains a challenging task, as traditional methods do not always adequately reflect the diversity and realism of the generated samples. This is particularly true for partial data generation, where changes are applied only to specific parts of an image, significantly complicating the assessment of their quality.*

This work examines various approaches to evaluating generative models, including automatic metrics such as Inception Score and Fréchet Inception Distance, precision, recall, density, and coverage, as well as a human-in-the-loop method such as HYPE. While these metrics have proven effective in evaluating the results of traditional generation, their use in the case of partially generated data may be inappropriate due to their limitations.

To address this issue, the paper proposes a new method for evaluating partially generated data that involves the human factor. This method is based on analysing transformed images by users, who identify the areas that have been altered, and evaluates their quality using precision, recall, and F1-score metrics by seeking intersections between actual altered areas and those selected by users using IoU. The proposed approach provides a more objective assessment of the realism and quality of generated image fragments during transformations.

A practical example of applying the developed method is presented using a dataset of panoramic dental images, where the quality of three models was evaluated: 1) a GAN based on a U-generator; 2) the same model with post-processing of the output image and segmentation mask; and 3) a self-validated GAN. The evaluation was performed by 30 individuals. The average F1-scores for these models were 0,78, 0,27, and 0,20, respectively. Since lower F1-scores in this case indicate better results (the more accurately users identified the transformations, the worse the model performed), the best model by this metric is the self-validated GAN, which is also supported by subjective assessments mentioned in the authors' work.

Keywords: *augmentation, data generation, generative adversarial network, GAN, computer vision, deep learning, self-validated GAN, evaluation, neural networks.*

ІНФОРМАЦІЯ ПРО АВТОРІВ

Ісаєнков Ярослав Олександрович — аспірант, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, Україна, 21021; +38 (095) 145-96-85; yisaienkov@gmail.com; ORCID: 0009-0005-5629-0021

Мокін Олександр Борисович — д-р техн. наук, проф., проф. кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, Україна, 21021; +38 (067) 785-98-44; abmokin@gmail.com; ORCID: 0000-0002-9277-3312

INFORMATION ABOUT THE AUTHORS

Isaienkov Ya. O. — Postgraduate Student, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, 95, Khmelnytsky Highway, Vinnytsia, Ukraine, 21021; +38 (095) 145-96-85; yisaienkov@gmail.com; ORCID: 0009-0005-5629-0021

Mokin O. B. — D. Sc. in Engineering, Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, 95, Khmelnytsky Highway, Vinnytsia, Ukraine, 21021; +38 (067) 785-98-44; abmokin@gmail.com; ORCID: 0000-0002-9277-3312