

## СИНТЕЗ ДАНИХ ЗА ДОПОМОГОЮ ГЕНЕРАТИВНИХ ЗМАГАЛЬНИХ МЕРЕЖ В СИСТЕМАХ ДЕТЕКЦІЇ ОБ'ЄКІВ

А. А. Пелих<sup>1, а</sup>, В. М. Ткач<sup>1</sup>, В. С. Левочко<sup>2</sup>

<sup>1</sup>Національний технічний університет України  
«Київський політехнічний інститут імені Ігоря Сікорського»,  
Фізико-технічний інститут

<sup>2</sup>Національний технічний університет України  
«Київський політехнічний інститут імені Ігоря Сікорського»,  
Інститут прикладного системного аналізу

### Анотація

Нещодавній прогрес у глибокому навчанні показав, що задачі комп'ютерного зору можна ефективно вирішувати за допомогою згорткових нейронних мереж. Проте, кількість наявних розмічених даних для тренування є однією з найважливіших проблем, що ускладнюють вирішення реальних завдань. Цю складність можна подолати, використовуючи синтетичні дані, отримані з різних джерел. У цій роботі представлений підхід до тренування мультирозмірного детектора об'єктів (Multi-scale Convolutional Neural Network, MS-CNN) на змішаному наборі даних. Показано, що доповнення тренувальної вибірки синтетичними зображеннями, отриманими за допомогою генеративних моделей машинного навчання, таких як генеративні змагальні мережі (Generative Adversarial Networks, GAN), може збільшити точність детекції пішоходів на наборі даних KITTI.

**Ключові слова:** глибоке навчання, комп'ютерний зір, генеративні моделі, генеративні змагальні мережі, синтетичні зображення, детекція об'єктів

### Вступ

Генерація зображень методами машинного навчання активно досліджується протягом останнього десятиліття. Проте, ніякого рішення, яке дозволяло б синтезувати реалістичні дані, не існувало до недавнього часу. Генеративні змагальні мережі (GAN), представлені Гудфелоу та ін. [1], значно сприяли успішному вирішенню цієї задачі. GAN складаються з двох змагальних моделей: генеративної моделі, яка імітує розподіл вхідних даних, та дискримінаційної моделі, яка оцінює ймовірність того, що елемент є реальним, а не штучно згенерованим. Мірза та Осіндеро [2] розширили GAN до умовної моделі, де генератор та дискримінатор приймають на вхід додаткову інформацію, таку як мітку класу або маску сегментації тощо. Дентон та ін. у своїй роботі [3] взяли за основу [1] та об'єднали умовну модель GAN з використанням представлення зображень у вигляді піраміди Лапласа, що призвело до підвищення якості згенерованих даних. Через деякий час Ізола та ін. [4] представили загальне рішення для трансляції зображення в зображення з використанням умовних GAN. Їх модель досягла значних результатів в широкому колі задач, таких як синтез фотографій з міток класів, реконструкція об'єктів з контурів та колоризація зображень. Причиною цьому стала покращена функція втрат, яка може адаптуватися до відмінностей між згенерованими та реальними зображеннями під час тренування.

Одним з останніх досягнень у синтезі зображень методами машинного навчання є модель Pix2PixHD, запропонована Вангом та ін. [5]. Їх метод мотивований успіхами Чена та Колтуна [6], які використали пряму регресійну ціль тренування, засновану на перцептивній функції втрат, та отримали першу модель, яка генерує зображення високої роздільної здатності 2048 x 1024 пікселів. Ванг та ін. [5] показали, що використання модифікованої цільової функції та мультирозмірних генераторів та дискримінаторів може не тільки стабілізувати тренування умовних GAN, але й досягти значно кращих результатів у порівнянні з Ченом і Колтуном [6].

### 1. Підходи до синтезу зображень

#### 1.1. Рендеринг комп'ютерної графіки

Прогрес у комп'ютерній графіці наразі дозволяє генерувати фотореалістичні віртуальні світи з автоматичною розміткою. Прикладом використання цього підходу є набір даних Virtual KITTI, запропонований в [7]. Хоча реальні дані з автоматичним розмічуванням і можуть бути отримані, проте створення сцени в програмному забезпеченні для рендерингу займає багато часу і потребує значного людського втручання.

#### 1.2. Комп'ютерні ігри

Іншим способом швидко отримувати візуальні дані та їх детальну семантичну розмітку є сучасні ком-

<sup>а</sup>anpel161@gmail.com



Рис. 1. Синтетичні зображення, отримані за допомогою Pix2PixHD, використовуючи маски семантичної сегментації Cityscapes

п'ютерні ігри. За допомогою цього підходу можна отримати велику кількість реалістичних зображень з різними сценами та об'єктами. Оскільки програмний код зазвичай недоступний, інформація з гри може бути здобута, використовуючи канали комунікації з графічним устаткуванням комп'ютера [8]. Проте, цей підхід вимагає додаткового програмного шару між грою та операційною системою, що дозволив би записувати, змінювати та відтворювати команди рендерингу. Такий інструмент створюється окремо для кожної гри, що потребує знання внутрішньої організації її компонентів та значних затрат часу.

### 1.3. Генеративні моделі

Враховуючи останні досягнення у сфері генеративних моделей машинного навчання, вони також можуть бути використані для синтезу реалістичних зображень. Після тренування вони дозволяють отримувати нові дані з мінімальним залученням людини. У даній роботі в якості основного підходу для генерації зображень використовується попередньо натренована модель Pix2PixHD [5].

В якості вхідних даних для моделі були використані карти семантичної сегментації з набору Cityscapes [9]. У результаті було отримано синтетичну версію Cityscapes, яка містить 3475 реалістичних зображень дорожніх сцен (приклад на рис. 1). Вони були використані для доповнення тренувального набору детектора об'єктів MS-CNN [10], який спочатку містив лише зображення з набору даних KITTI [11].

## 2. Результати

В якості тренувального набору для MS-CNN детектора було використано три комбінації даних:

А: тренувальний набір KITTI (7481 зображень);

В: тренувальний набір KITTI + 3475 згенерованих зображень;

С: тренувальний набір KITTI + 3475 зображень з Cityscapes;

Через використання MS-CNN великої кількості пам'яті та невеликий розмір міні-батчів, тренування проходило в два етапи для забезпечення стабільного багатозадачного процесу. На першому етапі тренується мережа регіональних пропозицій (Region Proposal Network) на всьому тренувальному наборі. Отримана модель використовується для ініціалізації другого етапу, що залучає підмережу детекції до тренувального процесу. Другий етап ділиться на дві фази: на першій фазі використовується тільки Cityscapes або згенеровані зображення, на другій – тільки KITTI зі зниженою швидкістю навчання.

Продуктивність MS-CNN детектора оцінювалася на даних KITTI [11] як одного з найбільш складних сучасних змагань з детекції об'єктів. KITTI містить три класи об'єктів: автомобілі, пішоходи та велосипедисти, та три рівні оцінювання: легкий, середній та важкий, що залежать від мінімальної висоти об'єкта та ступеня перекритості одного об'єкта іншими. Середній рівень складності найчастіше використовується для порівняння результатів. Якість роботи моделі оцінювалася за допомогою метрики інтерпольованої середньої точності (mean Average Precision, mAP), як описано в [12]. Отримані результати показані в табл. 1.

## Висновки

У роботі показано, що використання синтетичних зображень при тренуванні може збільшити точність детекції пішоходів на наборі даних KITTI. Було досягнуто збільшення середньої точності приблизно на 2% для всіх рівнів оцінки, порівняно з меншим набором реальних даних. Хоч приріст точності детекції і нижчий, ніж з використанням додаткових реальних зображень, даний експеримент демонструє потенціал використання синтетичних даних, отрима-

Табл. 1. Точність детекції на наборі даних KITTI, де А включає в себе тренувальний набір KITTI, В доповнений 3475 згенерованими зображеннями, С – суміш тренувальних наборів KITTI та Cityscapes

Тренувальний набір	Клас об'єктів	Легкий	Середній	Важкий
А	Пішоходи	83.69%	72.73%	65.86%
	Велосипедисти	81.20%	72.13%	63.51%
В	Пішоходи	<b>86.21%</b>	<b>74.74%</b>	<b>67.34%</b>
	Велосипедисти	83.21%	71.26%	62.62%
С	Пішоходи	86.32%	75.41%	71.27%
	Велосипедисти	82.24%	73.79%	63.10%

них за допомогою генеративних моделей глибокого навчання, для тренування детектора об'єктів.

В той самий час можна побачити, що точність детекції велосипедистів на середньому та важкому рівнях впала приблизно на 1% у порівнянні з тренуванням лише на справжніх даних. Також видно різкий стрибок точності детекції пішоходів на важкому рівні при використанні додаткових реальних даних у порівнянні з синтетичними. Ці явища зумовлені недостатньою якістю синтезу дрібних об'єктів сучасними генеративними моделями.

Варто зауважити, що набори реальних та синтетичних зображень мають різні розподіли ймовірностей, що призводить до більш різноманітного результуючого набору даних. Цей факт сприяє кращій узагальнюючій здатності моделі, проте може негативно впливати на точність роботи на конкретному наборі даних, такому як KITTI. Щоб збалансувати продуктивність моделі між тренувальним та цільовим доменами, слід в майбутньому дослідити можливість застосування методів передачі навчання (transfer learning) та адаптації доменів (domain adaptation).

## Перелік використаних джерел

1. Generative adversarial networks. / Goodfellow, J. Pouget-Abadie, M. Mirza et al. // Advances in Neural Information Processing Systems (NIPS). — 2014.
2. Mirza M., Osindero S. Conditional generative adversarial nets // CoRR, abs. — 2014. — Vol. 1411.1784.
3. Deep generative image models using a laplacian-pyramid of adversarial networks / E. Denton, S. Chintala, R. Fergus, A. Szlam // Advances in neural information processing systems. — 2015.
4. Mirza M., Osindero S. Conditional generative adversarial nets // CoRR, abs. — 2014. — P. 1486–1494.
5. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs / T.-C. Wang, M.-Y. Liu, J.-Y. Zhu et al. // arXiv preprint arXiv:1711.11585.
6. Chen Q., Koltun V. Photographic image synthesis with cascaded refinement networks // IEEE International Conference on Computer Vision (ICCV). — 2017.
7. Virtual Worlds as Proxy for Multi-Object Tracking Analysis / A. Gaidon, Q. Wang, Y. Cabon, E. Vig // arXiv:1605.06457v1 [cs.CV].
8. Playing for Data: Ground Truth from Computer Games / S. Richter, V. Vineet, S. Roth, V. Koltun // 14th European Conference on Computer Vision (ECCV). — 2016.
9. The Cityscapes Dataset for Semantic Urban Scene Understanding / M. Cordts, M. Omran, S. Ramos et al. // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2016.
10. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection / Z. Cai, Q. Fan, R. S. Feris, N. Vasconcelos // European Conference on Computer Vision (ECCV). — 2016.
11. Vision meets Robotics: The KITTI Dataset / A. Geiger, P. Lenz, C. Stiller, R. Urtasun // International Journal of Robotics Research (IJRR). — 2013.
12. Geiger A., Lenz P., Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2012.