

Zobrazowania satelitarne: dane syntetyczne motorem szerszego wykorzystania sztucznej inteligencji

Za dużo i za mało

Do trenowania algorytmów AI potrzeba ogromnych zbiorów danych, których przygotowanie jest niezwykle pracochłonne. Prawdziwe zobrazowania można jednak zastąpić syntetycznymi, znacznie łatwiejszymi do pozyskania, a dającymi porównywalne wyniki.

Paweł Ziemnicki

Analitycy są obecnie wręcz zalewani strumieniem danych pochodzących z satelitarnej obserwacji Ziemi. Dopiero jednak przetworzenie i interpretacja tych danych pozwalają na wyciągnięcie z nich cennych informacji, ukazując ich rzeczywistą wartość dla użytkownika końcowego. W związku z tym oczywista jest konieczność przyspieszenia oraz automatyzacji pracy z danymi. Ale algorytmy, które mogą nam to zapewnić, najpierw same po-

trzebują się „wykształcić”, analizując duże zbiory danych uczących. Te właśnie kwestie były jednym z ciekawszych tematów poruszanych podczas GEOINT Symposium, które odbyło się w kwietniu w stanie Kolorado. To największe zgromadzenie specjalistów zajmujących się wywiadem geoprzestrzennym organizowane jest w Stanach Zjednoczonych co roku. Na wydarzenie skadają się m.in. przemówienia liderów, spotkania tematyczne, profesjonalne szkolenia i sesje edukacyjne oraz wystawy.

W czasie tegorocznego sympozjum prelegenci przekonywali, że choć zbiory zo-

brazowań satelitarnych dostępne dziś dla naukowców, analityków czy przedstawicieli wywiadów wojskowych bezustannie się rozrastają, to niektóre typy obiektów wciąż bardzo rzadko się na nich pojawiają. Chodzi tu na przykład o specyficzne modele okrętów podwodnych, wyrzutni raketowych czy bombowców. Jeśli dysponujemy jedynie niewielkim zbiorem zdjęć satelitarnych danego obiektu, nie jesteśmy w stanie zbudować wystarczająco dużej bazy danych do wyuczenia algorytmów. W efekcie nie będziemy mogli przygotować narzędzi opartych na sztucznej inteligencji



Przykład tworzenia danych syntetycznych dla branży teledetekcyjnej



Źródło: Intelligence-tribusfs.com

©Oneview

Detekcja konkretnych typów statków powietrznych na różnym tle

(Artificial Intelligence – AI) czy też uczeniu maszynowym (Machine Learning – ML), które mogłyby w przyszłości automatycznie rozpoznawać interesujące nas obiekty na nowo pozyskiwanych zobrazowaniach.

• Niewystarczająca ilość danych

Corey Jaskolski, założyciel i prezes firmy Synthetiaic, w marcu 2021 r. tak pisał o problemie na portalu Forbes.com: „To powszechnie panujące błędne przekonanie, że większość firm tonie w danych, a sztuczna inteligencja ma niemal nieograniczony wybór, jeśli chodzi o dane

szkoleniowe. Prawda jest taka, że mimo boomu na duże zbiory danych, większości firm wciąż brakuje odpowiedniej ilości danych wysokiej jakości, co wstrzymuje rozwój najbardziej wartościowych aplikacji AI”.

Przedsiębiorstwo L3Harris również podkreśla na swojej stronie internetowej, że postęp w dziedzinie AI jest skoncentrowany wokół danych, których paradoksalnie jest jednocześnie za dużo i za mało. Modele wykorzystujące uczenie maszynowe będą działać prawidłowo dopiero po „przetrawieniu” znacznej ilości danych (w tym przypadku zdjęć) dotyczących obiektu, który mają wykrywać.

– Mówiąc o „wykrywaniu”, odnosimy się do czynności polegającej na identyfikacji różnych obiektów na obrazach satelitarnych w celu zbierania danych wywiadowczych. Gdy potrzebne jest wykrywanie obiektów na dużą skalę, na przykład przedzieranie się przez tysiące czy miliony obrazów, stosowane są algorytmy ML – objaśniają z kolei specjaliści z firmy OneView.

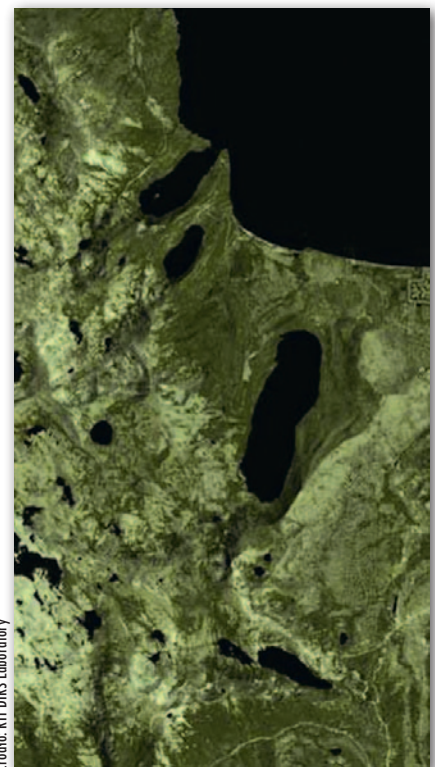
Podczas kwietniowego GEOINT w tym właśnie kontekście kierujący firmą Orbital Insight Kevin O’Brien zapowiadał, że przyszłością są dane syntetyczne, czyli sztucznie kreowane zobrazowania satelitarne. Wykorzystanie tego typu danych referencyjnych do ćwiczenia algorytmów w wykrywaniu rzadko spotykanych obiektów może być przydatne nie tylko z wojskowego punktu widzenia, np. dla odbiorców takich jak amerykańska NGA (National Geospatial-Intelligence Agency, Narodowa Agencja Wywiadu Geoprzestrzennego), ale również dla szeroko rozumianego przemysłu. Orbital Insight pracuje na zlecenie NGA nad potencjalnym wykorzystaniem syntetycznie generowanych danych do

trenowania sztucznej inteligencji w kierunku wykrywania nowych obiektów na zobrazowaniach satelitarnych RGB.

Inny gracz obecny na GEOINT 2022, wspomniany już L3Harris, pochwalił się wypracowanymi przez swoich inżynierów sposobami przygotowania syntetycznych danych treningowych zarówno w zakresie zobrazowań RGB, jak i radarowych SAR. – Skoncentrowaliśmy się na dokładności danych syntetycznych. Tylko kiedy zachowujesz tę dokładność, szkolenie algorytmów AI może być porządne – mówiła Stacey Casella, starsza dyrektor ds. przetwarzania danych



Źródło: OneView



Źródło: RTI DIRSIG Laboratory

Przykład jednokanałowego obrazu satelitarnego utworzonego z użyciem DIRSIG



Źródło: CosmIQ Works i Al.Reverie

Obrazy z zestawu danych wysokorozdzielczych „RarePlanes” opublikowanego w 2020 r. Dwa górne rzędy zawierają zobrazowania z satelity WorldView-3 Maxar Technologies, a dwa dolne – dane syntetyczne

i analiz geoprzestrzennych w L3Harris. Fachowcy z tego przedsiębiorstwa demonstrowali, jak produkują syntetyczne obrazy myśliwców umieszczonych na różnych płach, przy różnych warunkach pogodowych z uwzględnieniem różnych kątów patrzenia sensora czy padania promieni słonecznych.

● Wąskie gardło

Dino Mintas, główny specjalista naukowy ds. danych w VISIMO, potwierdza, że do trenowania modeli ML niezbędne są ogromne zbiory danych, a niedobór ich jest często czynnikiem hamującym rozwój algorytmów. Tworzenie odpowiednich zbiorów danych uczących wymaga niejednokrotnie zaangażowa-

nia potężnych zasobów, na co nie każdy przedsiębiorca może sobie pozwolić. Po pierwsze, trzeba zgromadzić znaczną ilość danych. Po drugie, dane te muszą spełniać kryterium różnorodności, tak by jak najszerzej odzwierciedlały warunki występujące w różnych częściach świata. Poszczególne zobrazowania tego samego obiektu mogą się bowiem znacznie różnić. Wymieńmy tu tylko kilka zmiennych dla przykładu: kąt obserwacji, pora dnia czy warunki pogodowe. Katalog obrazów referencyjnych o odpowiednim zróżnicowaniu nie zawsze jest dostępny.

Ponadto zbieranie danych treningowych może być procesem kosztownym i długotrwałym. Będzie to odczuwalne szczególnie przy mozolnym ręcznym

tworzeniu odpowiedniej bazy przez odpowiednio wykwalifikowanych ekspertów, zmuszonych poszukiwać rzadkich, wysoce specjalistycznych obrazów. Następnie zdjęcia takie muszą być kategoryzowane oraz opatrywane odpowiednimi adnotacjami, co również jest zadaniem dla fachowców.

Problematyka ta jest rozwinięta na stronie firmowej One-view.ai: „Zbieranie i przygotowywanie rzeczywistych danych stanowi wąskie gardło spowalniające cały proces analizy obrazów opartej na ML. Główną przeszkodą jest ręczne dodawanie adnotacji do zbioru danych, co w sposób oczywisty jest czasochłonne, podatne na błędy i wymaga wielu czynności (...), aby uzyskać właściwe adnotacje. Dotyczy to szczególnie zdjęć satelitarnych, na których obiekty są trudne do zauważenia, stosunkowo małe i o niskiej rozdzielczości”.

W październiku 2021 r. na portalu Techcrunch.com zwracano uwagę, że wyżej wymienione trudności mogą być jeszcze większe w przypadku obrazów, które nie są typowymi zobrazowaniami RGB, a na przykład radarowymi. Natomiast na ograniczenia związane z koniecznością wykorzystania wyspecjalizowanych zasobów ludzkich zwrócili uwagę fachowcy z L3Harris. Wskazali, że algorytmy uczenia maszynowego pozwalają szybko identyfikować konkretne obiekty w olbrzymich zbiorach danych, ale dla swego prawidłowego działania wymagają od analityków nakreślenia odpowiednich ram kontekstowych, na podstawie których przebiegać będzie proces uczenia tychże algorytmów. Usprawnienie tego procesu



Przykład oznakowanych danych syntetycznych

Źródło: L3Harris

pozwole na tworzenie wysoce wyrafinowanych platform opartych na AI.

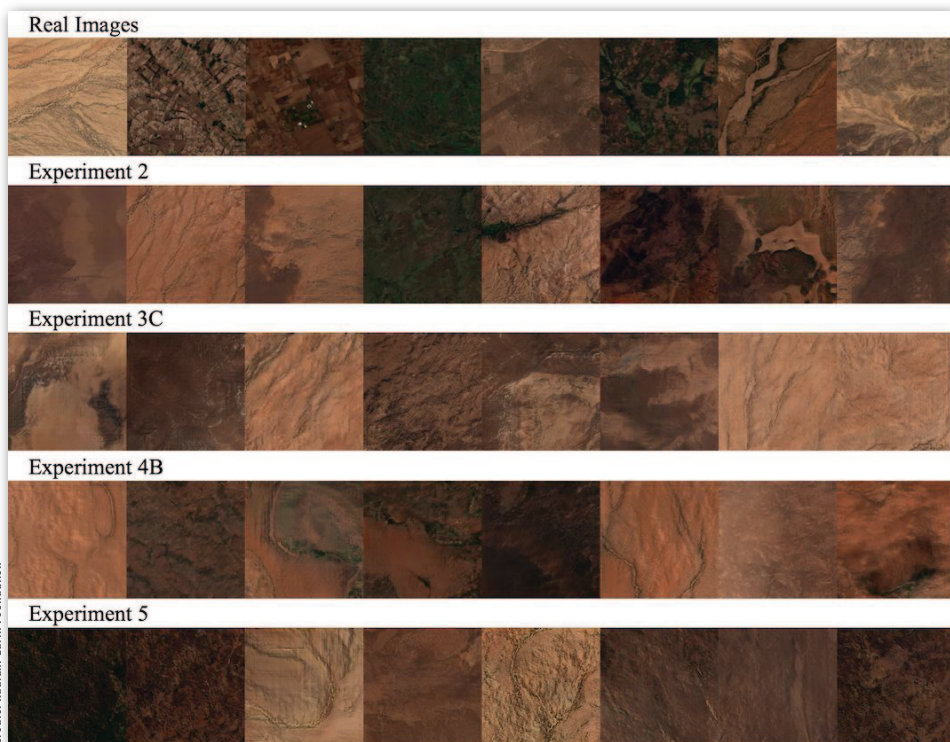
• Kluczem dane syntetyczne

Problem braku dostępu do danych bądź też ich ograniczonej ilości może być dziś rozwiązany dzięki tworzeniu danych syntetycznych w postaci sztucznych zobrażeń satelitarnych. Takie dane tworzy się od podstaw. Można je bardzo precyzyjnie dostosować do konkretnych potrzeb treningowych, sterując dziesiątkami parametrów. To gwarantuje otrzymanie niezbędnej różnorodności w tworzonych zbiorach. Dane syntetyczne powstają dość szybko, a właściwe adnotacje są w nich „zaszywane” już na etapie tworzenia. W efekcie otrzymujemy odpowiednio oznaczony zestaw danych obrazowych gotowy do trenowania modeli ML.

– Kiedy używamy terminu dane syntetyczne, mamy na myśli zaprojektowane metodami inżynierskimi symulowane zbiory danych, a szczególnie skupiamy się na symulacji opartej na fizyce – mówi dyrektor Nathan Kundtz, CEO w firmie Rendered.ai. Są to dane sztucznie wytwarzane, a nie zbierane w świecie rzeczywistym. – Dzięki wykorzystaniu obrazów generowanych komputerowo, które ściśle odpowiadają obrazom ze świata rzeczywistego, szybko tworzone są zestawy danych treningowych ze znacznymi wariacjami. Dane syntetyczne zapewniają skalowalną metodę generowania danych treningowych bez konieczności poświęcania znacznych zasobów na ręczne zbieranie, etykietowanie i selekcjonowanie danych uczących na podstawie rzeczywistych obrazów. Ta technologia zmniejsza koszty, ryzyka oraz prawdopodobieństwo stronnego doboru danych i umożliwia bardziej dokładne, godne zaufania działanie sztucznej inteligencji – podsumowują zaopiniowanie specjaliści z L3Harris

Zdaniem ekspertów nie jest możliwe, żeby dane syntetyczne w pełni zastąpiły dane rzeczywiste. Mogą jednak znacznie przyspieszyć rozwój modeli wykorzystujących mechanizmy sztucznej inteligencji, a ich rola będzie jeszcze rosła. Nathan Kundtz uważa, że rozwój technologii w zakresie tworzenia sztucznych danych może w pewnym zakresie podważyć monopolistyczną pozycję gigantów, takich jak Google, dysponujących bilionami zobrażeń oraz nieprzebranym ogromem zbiorów danych.

W kwietniu 2022 roku opublikowano informację o nawiązaniu przez Rendered.ai współpracy z RIT (Rochester Institute of Technology) Digital Imaging and Remote Sensing Laboratory (DIRS). W ra-

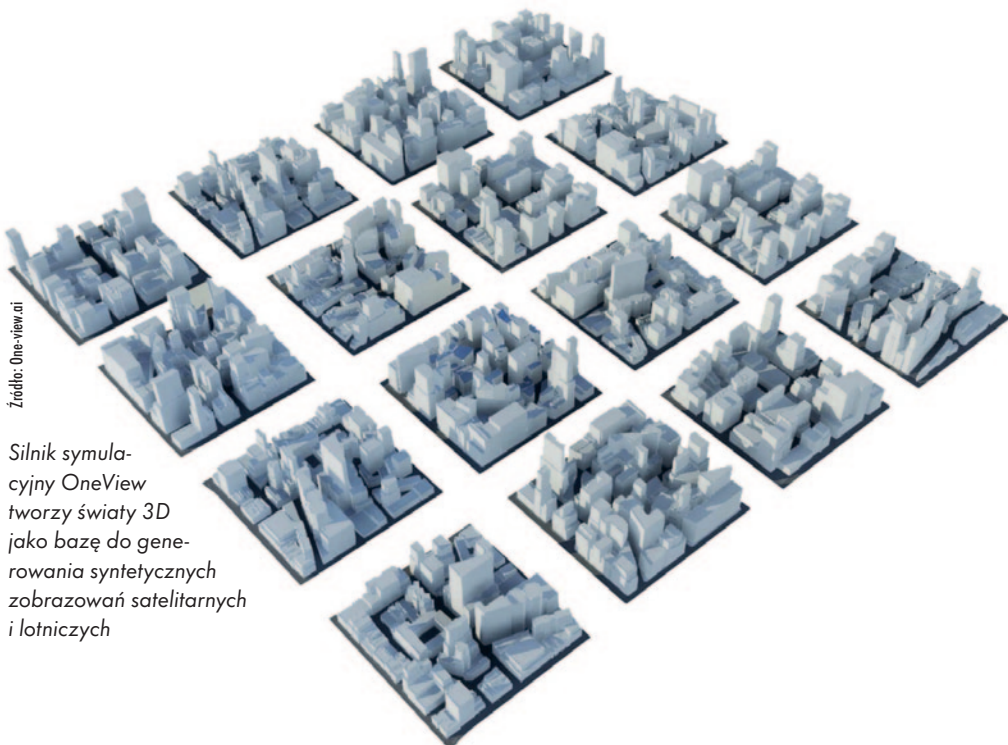


Efekty eksperymentów z przygotowaniem zobrażeń syntetycznych w pasmach R, G, B

mach kooperacji Rendered.ai wnosi platformę chmurową do tworzenia danych syntetycznych na masową skalę. DIRS Laboratory z RIT udostępnia natomiast DIRSIG, czyli model Digital Imaging and Remote Sensing Image Generation. Jest to oparty na prawach fizyki model do generowania obrazów syntetycznych.

Tworzenie obrazów syntetycznych w zgodzie z prawami fizyki jest ważne i wymaga więcej wysiłku niż przykładowo proste przeniesienie wizerunku pojazdu opancerzonego z tła nawierzchni asfaltowej na tło nawierzchni trawiastej. W tym modelu uwzględnia się szereg parametrów obserwacyjnych, takich jak

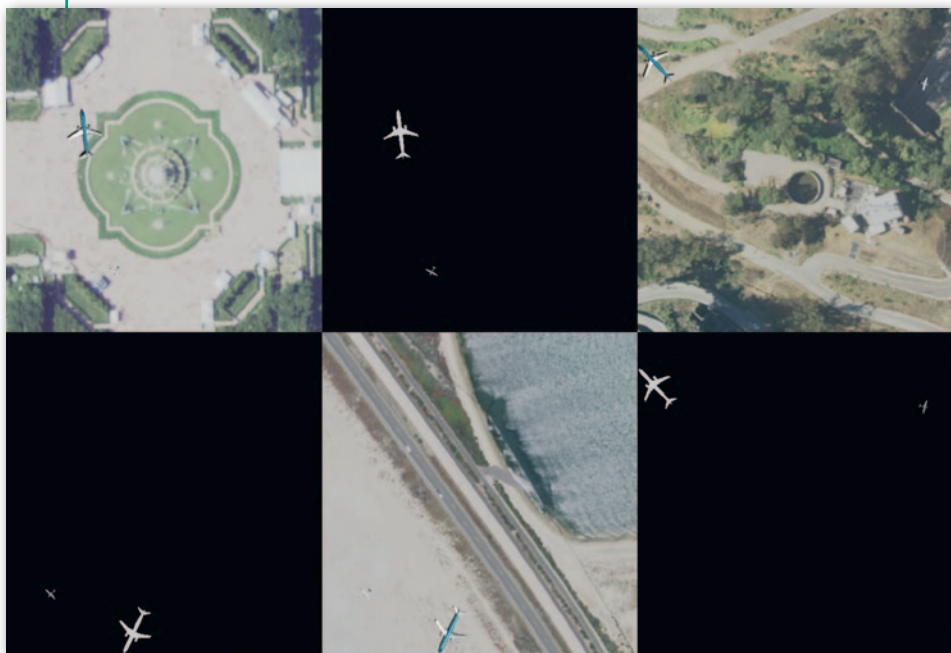
kąt padania promieni słonecznych czy wilgotność powietrza. Sztucznie generowany obraz ma możliwie najwierniej odzwierciedlać to, w jaki sposób te symulowane warunki wpłynęłyby na jasność, kolor czy ostrość obserwowanego obiektu na zdjęciu. DIRSIG umożliwia tworzenie pasywnych obrazów jednokanałowych, multispektralnych czy nawet hiperspektralnych, pokrywając zakresy od światła widzialnego do obszaru podczerwieni termalnej widma promieniowania elektromagnetycznego. Model ten może służyć do testowania projektów systemów obrazowania, generowania obrazów testowych do oceny algorytmów



Silnik symulacyjny OneView tworzy światy 3D jako bazę do generowania syntetycznych zobrażeń satelitarnych i lotniczych

Źródło: Rendered Earth Foundation

Źródło: One-view.ai



Źródło: Rendered.ai

Obraz z Rendered.ai, który łączy usługi obrazowania ArcGIS i modele 3D samolotów. Obrazy przedstawiają wygenerowane syntetyczne zobrazowania i dokładne maski, które precyzyjnie informują algorytm, gdzie znajdują się samoloty

przetwarzania obrazów oraz do tworzenia danych do szkolenia analityków pracujących z danymi obrazowymi.

– DIRSIG od lat dostarcza syntetyczne obrazy wyspecjalizowanym klientom – tłumaczy dr Scott Brown, główny naukowiec i kierownik projektu. – Współpraca z Rendered.ai umożliwia udostępnienie naszych sprawdzonych zdolności szerszemu gronu odbiorców w czasie, kiedy branża danych satelitarnych i inne formy gromadzenia danych z teledetekcji szybko się rozwijają – podkreśla.

• Kwestia dni lub nawet godzin

Ekspertcy są zgodni, że algorytmy uczenia maszynowego stanowią dziś klucz do maksymalnego wykorzystania możliwości, jakie w skali globalnej są cały czas szybko rozwijane przez podmioty zajmujące się zbieraniem i analizą zdjęć satelitarnych. – Ekspansja rynku obserwacji Ziemi nie tylko zwiększa różnorodność typów dostępnych danych i częstotliwość ich gromadzenia dla użytkowników rządowych i komercyjnych, ale także napędza potrzebę znalezienia nowych i szybszych zautomatyzowanych technik wydobywania wiedzy z ogromnych strumieni zobrazowań – tłumaczy Nathan Kundtz. Produkcja danych syntetycznych to przełomowa technologia, dzięki której możliwe jest pokonanie problemu wąskiego gardła, jakie niedobór danych treningowych może obecnie stanowić dla rozwoju rozwiązań opartych na AI oraz ML.

Cytowany już Corey Jaskolski ocenia, że chociaż syntetyczne zobrazowania sa-

telitarne należą do najtrudniejszych tego typu danych do wygenerowania, to jednocześnie mogą się zaliczać do najbardziej wartościowych. W swoim artykule na portalu Forbes.com patrzy w przyszłość tej dziedziny z nieskrywanym optymizmem. – Bez wymaganych danych odpowiedniej jakości nie byliśmy w stanie zbudować dokładnych, predykcyjnych modeli sztucznej inteligencji tam, gdzie są one najbardziej potrzebne. To się wkrótce zmieni. Dysponując technologią danych syntetycznych, odpowiednie dane biznesowe (których zebranie w innym przypadku trwałoby miesiące czy lata lub które są po prostu niedostępne) będziemy mogli „wychodować” w ciągu dni, a w niektórych przypadkach zaledwie godzin – przewiduje specjalista.

Firma Mostly AI, która zajmuje się danymi syntetycznymi różnego rodzaju (m.in. dla telekomunikacji, ubezpieczeń czy bankowości) przekonuje na swojej stronie, że „dane syntetyczne oferują prawdziwą reprezentację świata rzeczywistego”, a zawarte w nich „wzorce, korelacje i fragmenty informacji pozostają takie same”. Wydaje się, że rozwiązanie to sprawdza się także w przypadku trenowania algorytmów pod kątem detekcji obiektów na podstawie syntetycznych danych teledetekcyjnych.

Autorzy artykułu „Generating Synthetic Data for Remote Sensing”, który w 2020 r. ukazał się na portalu Old.trajectorymagazine.com, zapewniają, że przy zastosowaniu odpowiedniej metodyki da się osiągnąć taki efekt,

kiedy to „z jakościowego punktu widzenia trudno jest odróżnić rzeczywiste od syntetycznych zbiorów danych”. Natomiast w kwietniu br. Eric van Rees zapewnił w swoim tekście dla „Geo Week News”, że dane syntetyczne z pewnością mogą reprezentować istniejące lokalizacje geograficzne na planecie. W tym kierunku idą zresztą już całkiem poważni gracze w branży. Zaraz potem Rees dodaje bowiem, że w ostatnim czasie firma Rendered.ai dołączyła do programu Esri jako partner startupowy na rzecz wykorzystania danych GIS i zobrazowań jako tła i konteksty dla środowisk syntetycznych.

Interesujące i optymistyczne są ponadto wyniki przeprowadzonego przez specjalistów z OneView studium przypadku, którymi firma dzieli się na swojej stronie internetowej: „Wyniki zachęciły nas do zbadania, czy możemy osiągnąć dalszą poprawę, łącząc zarówno rzeczywiste, jak i syntetyczny zestaw danych. (...) To podejście okazuje się bardzo skuteczne, (...) nawet symboliczne dodanie obrazów rzeczywistych znacznie poprawia wyniki, pokazując, że niewielka ilość obrazów rzeczywistych i duża ilość danych syntetycznych daje w efekcie wysokiej jakości algorytm”.

Zaangażowani w badanie eksperci tak podsumowują jego rezultaty: „Wykazaliśmy, że gdy wielkość próbki jest niewielka, dane syntetyczne sprawdzają się znacznie lepiej niż dane rzeczywiste, co skutkuje wysokiej jakości algorytmem – osiąganym w krótszym czasie, mniejszym nakładem pracy i przy niższym koszcie. Dodatkowo wykazaliśmy, że łączenie danych rzeczywistych i syntetycznych do trenowania modeli daje najlepsze wyniki. Idealny »miks« to niewielka liczba rzeczywistych zobrazowań z dużą ilością danych syntetycznych”.

Paweł Ziemiński

dziennikarz, ekspert ds. tematyki kosmicznej, pracuje na rzecz podmiotów związanych z satelitarną obserwacją Ziemi

SPROSTOWANIE

W opublikowanym w GEODECIE 5/2022 wywiadzie z prof. Katarzyną Osińską-Skotak ukazała się nieprecyzyjna informacja dotycząca danych wykorzystywanych przez firmę Tensorflight do detekcji pokryć dachowych. Obok danych satelitarnych firma posługuje się danymi z mobilnego skanowania laserowego (MLS), a nie konkretnie danymi Street View. Za nieścisłość przepraszamy Czytelników i firmę Tensorflight.

Redakcja