DOI: https://doi.org/10.15276/ict.02.2025.20
УДК: 004.93

Адаптивний фрагментний аналіз відеопотоків для класифікації дорожньо-транспортних пригод за допомогою розрідженого відео-трансформеру

Норматова Тетяна Віталіївна[1)]
Аспірантка каф. Інформатики
ORCID: https://orcid.org/0009-0004-3503-6350, tetiana.normatova@nure.ua
Машталір Сергій Володимирович[1)]
Д-р техніч. наук, професор каф. Інформатики
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Харківський національний університет радіоелектроніки, пр. Науки, 14. Харків, 61166, Україна

АНОТАЦІЯ

У роботі запропоновано простий у програмній реалізації та дієвий підхід до класифікації коротких відеофрагментів на аварійні та нормальні сцени. Для даної роботи використовується рівномірна вибірка кадрів з усього кліпу (6–8 кадрів) для того, щоб не втрачати ключові події навіть у довгих роликах. Далі застосовується адаптивна «фрагментація» кадру, керована рухом: за картою оптичного потоку Farnebäck обчислюються квантільні пороги й для кожної комірки базової сітки обирається розмір фрагмента (8/16/32 пікселів). У ділянках із виразним рухом обираються дрібні фрагменти (вища деталізація), у статичних — більші (менше обчислень). Відібрані фрагменти не перекриваються, масштабуються до базового розміру та перетворюються на ознакові вектори. Архітектура побудована за двоступеневим принципом. На першому кроці просторовий блок уваги працює всередині одного кадру — лише над відібраними фрагментами, що суттєво зменшує кількість ознакових одиниць. На другому кроці часовий блок опрацьовує послідовність кадрів через їх короткі підсумкові представлення (службові класифікаційні маркери, далі — CLS), агрегуючи динаміку у часі. Така факторизація «простір → час» знижує обчислювальну вартість і пам'ять без втрати інформативності в рухомих регіонах. Для боротьби з дисбалансом класів застосовано зважену функцію втрат (або «втрату з фокусуванням на важких прикладах») і випадкове вибіркування з вагами під час навчання. Попередньо на диск зберігаються карти оптичного потоку та списки відібраних фрагментів, що прискорює епохи на процесорі без спеціального графічного обладнання. Оцінювання проводиться на CCD1500 (1500 аварійних і 3000 нормальних відео) зі стандартним поділом 80/20 за збереженням часток класів. Отримано точність 0.864 і макро-F1 0.851; за попереднім порівнянням запропонований підхід перевершує базову рівномірну розбивку кадру та класичні схеми з простим часовим вибіркуванням. Головна цінність підходу — поєднання «рух-керованого» скорочення ознакових одиниць і двоступеневої обробки, що робить модель придатною для реалістичних обмежень за часом і ресурсами (CPU), зберігаючи високу чутливість до коротких і локальних аварійних подій. Метод легко масштабувати та поєднувати з попереднім навчанням на основі маскованих відновлень відео. Також описується фіксацію умов, відкриті налаштування й кроки для повної відтворюваності.

Ключові слова: відеокласифікація; нейронні мережі; згорткові нейронні мережі; класифікація об'єктів; аналіз відеопотоків, класифікація даних; обробка фрагментів зображення

Штучний інтелект сьогодні широко застосовується для аналізу дорожнього руху й виявлення аварійних ситуацій. Більшість підходів спирається на 3D-згорткові нейронні мережі, що обробляють весь кадр послідовно. Однак висока розмірність відеоданих призводить до значних обчислювальних затрат і значної кількості параметрів. У даній роботі пропонується зменшити непотрібні обчислення шляхом селективного виділення фрагментів із кадру: регіони з великою швидкістю руху описуються дрібними фрагментами, а з низькою швидкістю – більшими. Такий підхід дозволяє сфокусувати увагу мережі на динамічних областях, що найважливіші для виявлення аварій.

До архітектури описаного методу входить двоступеневий Sparse-ViT: просторовий блок уваги працює над фрагментами зображення кожного кадру, а часовий блок – над послідовністю службових токенів-класифікаторів (CLS), які узагальнюють інформацію по кадрах. Перед векторизацією виконується плиткова (без перекриття) адаптивна фрагментація, керована оптичним потоком: для кожної базової комірки оцінюється модуль вектору потоку відносно квантильних порогів (q_1/q_2) і обирається розмір фрагмента (наприклад, 8/16/32 пікселів). Фрагменти вирізаються без перекриття, масштабуються до базового розміру та перетворюються на векторні ознаки. Такий відбір зменшує кількість зайвих ознак і спрямовує обчислення на ділянки з рухом, що підвищує результативність як просторового аналізу кадру, так і узгодження інформації в часі.

Transformer-підходи до відео. ViT [1] заклав основу для подання зображень як послідовності невеликих фрагментів, кожен з яких перетворюється на вектор ознак; TimeSformer [2] розділяє увагу на просторову й часову; ViViT [3] факторизує просторово-часову увагу, застосовуючи окремі блоки для фрагментів кадру та агрегації по часу; Video Swin Transformer [4] вводить ієрархічні зсувні вікна; MVi [5] формує багаторівневі подання. Ці лінії робіт демонструють, що розділення простору й часу та багатомасштабність покращують ефективність, але квадратична обчислювальна вартість механізму повної уваги над усією послідовністю ознак залишається вузьким місцем.

Розріджена/ефективна увага та довгі послідовності. Моделі на кшталт Sparse Transformer**,** Longformer і BigBird зменшують обчислювальну складність завдяки локальним та випадковим схемам зосередження уваги; вибіркова увага зі зміщеннями показала, що відбір невеликої кількості найважливіших опорних елементів може замінити повну глобальну увагу. Для відео це особливо важливо під час обробки довгих відрізків. У підході, який описано у даній статті, «розрідженість» досягається не шляхом зміни механізму внутрішньої уваги, а завдяки попередньому скороченню кількості елементів послідовності: залишаються лише фрагменти кадру з підвищеною ознакою руху, тож довжина послідовності зменшується ще до етапу обробки перетворювачем**.**

Оптичний потік і двопотокові сигнали. FlowNet2, PWC-Net та RAFT зробили потік надійним джерелом ознак руху. Рух корисний і як окремий вхід (two-stream), і як підказка для просторового відбору та приховування ознакових елементів. У описаній реалізації використовується класичний Farnebäck-flow у якості «карти важливості», що керує адаптивним патчингом; це дає значний виграш у швидкості на CPU при прийнятній якості.

Відбір і скорочення подань. Щоб не опрацьовувати всі фрагменти кадру, метод TokenLearner динамічно об’єднує невелику підмножину найінформативніших фрагментів; DynamicViT навчається пропускати малозначущі фрагменти під час застосування моделі. Для відео з’явилися підходи, керовані рухом (наприклад, варіанти ViViT/MotionFormer), а також методи зосередження на важливих фрагментах (AdaFocus), які підвищують роздільність у суттєвих ділянках. У даній статті дотримується ця лінія: застосовується адаптивний відбір фрагментів, де розмір вікна огляду (8/16/32 пікселів) визначається за квантильними порогами величини оптичного потоку; фрагменти вирізаються без накладання і приводяться до базового розміру, що помітно зменшує кількість опрацьовуваних подань, зберігаючи зони з вираженим рухом.

Дисбаланс класів. Поширені підходи – збалансована крос-ентропійна втрата або повторне зважування**;** *Focal Loss* – спеціалізована функція втрат [6]; надвибірка через зважений випадковий добір (WeightedRandomSampler). Для відео також застосовують сегментне вибіркове формування кадрів методом часової сегментації (Temporal Segment Networks, TSN), щоб не «тонути» в довгих нормальних фрагментах. У описаній реалізації вже використано повторне зважування у функції втрат і зважений випадковий добір у завантажувачі даних; Focal Loss та TSN-стиль сегментування можна ввімкнути за потреби.

Попереднє навчання та масковане моделювання (перспектива). MAE і VideoMAE показують, що приховування фрагментів із подальшим відновленням суттєво підвищує стійкість ознак, особливо коли розмітка даних обмежена. У даній роботі поки що не застосовувався підхід MAE у наведених експериментах, однак він сумісний із модулем формування та відбору фрагментів кадру, що використано у даній роботі, і є логічним наступним кроком.

Підсумок позиціонування. На відміну від підходів, що вдосконалюють саму схему уваги у поточній роботі знижено обчислювальну вартість завдяки відбору фрагментів за показником руху та двоетапній обробці: просторовий етап працює з відібраними фрагментами кожного кадру, а часовий – узагальнює лише службові підсумкові позначки кадрів (CLS). Така схема дає компактну, придатну для звичайного процесора конфігурацію без надмірно складних складників і з помірними вимогами до ресурсів та обсягу даних.

Адаптивний вибір фрагментів (motion-guided)

Вхід і відбір фрагментів. $Xt \in \mathbb{R}^{H\times W\times 3}$ кадр , карта величини потоку M_tміж X_t *і* X_{t+1}. Пороги q_1, q_2 обчислюються по ненульових значеннях M_t (33% і 66%). Базова сітка з кроком b (типово b = 8); кількість базових комірок на кадр:

$$N_{base} = \left\lfloor \frac{H}{b} \right\rfloor \cdot \left\lfloor \frac{W}{b} \right\rfloor,$$

решта краю кадру відсікається.

Для кожної комірки обчислюється стійка оцінка інтенсивності руху

$$\bar{M}_c = \frac{1}{k}\sum_{(x,y)\in Top-k(c)} M_t(x,y), \; k = \lfloor 0.25b2 \rfloor$$

Вибір масштабу фрагменту:

$$\begin{cases} \bar{M}_c \geq q_2 \Rightarrow s_c = b, \\ q_1 \leq \bar{M}_c < q_2 \Rightarrow s_c = 2b, \\ \bar{M}_c < q_1 \Rightarrow s_c = 4b. \end{cases}$$

Без перекриття: розбиття виконується лише в межах комірок із призначеним розміром s_c. Реалізація**:** ієрархічний поділ на чверті або просте об’єднання сусідніх комірок з однаковим s_c. Після вирізання кожен фрагмент змінюється до розміру $b\times b$. Вводиться верхня межа кількості фрагментів на кадр K_{max}; якщо фрагментів більше – залишаються перші K_{max} за значенням $\bar{M}_c$; якщо жодної комірки не відібрано – повертається рівномірна сітки $b\times b$. На практиці кількість фрагментів зменшується приблизно на 30-60 % порівняно з рівномірною сіткою (залежно від сцени). Результат роботи алгоритму адаптивного поділу на фрагменти можна побачити на рис.

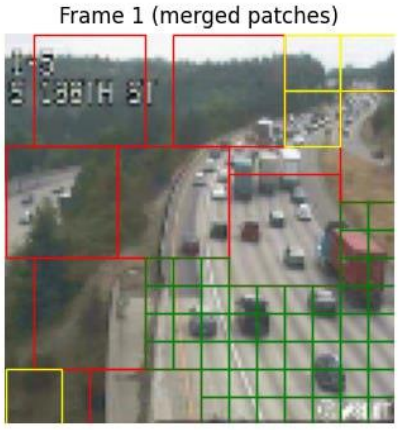

Рисунок. **Візуалізація адаптивного відбору фрагментів: зелений – базові фрагменти *b=8×8*; жовтий – збільшені *2b=16×16*; червоний – великі *4b=32×32*. Дрібні фрагменти розміщуються в ділянках активного руху (смуги з транспортом), великі – на статичних областях фону**

Двоступеневий розріджений відео-трансформер

Просторовий блок. Застосовується механізм багатоголової уваги (multi-head attention) над ознаками фрагментів одного кадру. Попередньо кожен фрагмент розміром $b\times b$ перетворюється у вектор $x_i = vec(P_i) \in \mathbb{R}^{3\,b^2}$ і проектується в простір розмірності d:

Для кожного вибраного фрагмента розміру $b\times b$ формується вектор $x_i \in \mathbb{R}^{3b^2}$ і виконується вбудовування фрагментів (patch-embedding):

$$z_i^0 = E\,x_i + e_i^{pos},\; E \in \mathbb{R}^{d\times 3\,b^2},\; e_i^{pos} \in \mathbb{R}^d,$$

де $P_i \in \mathbb{R}^{b \times b \times 3}$ – фрагмент кадру; $x_i = vec(P_i) \in \mathbb{R}^{3\,b^2}$ – його векторизація; $E \in \mathbb{R}^{d \times 3\,b^2}$ – матриця лінійної проєкції (спільна для всіх фрагментів); $e_i^{pos} \in \mathbb{R}^d$ – позиційне вбудовування, а $z_i^0 \in \mathbb{R}^d$ – ознаковий вектор фрагмента на вході першого шару перетворювача

Далі формується послідовність з доданим службовим маркером [CLS]:

$$Z^0 = [\, z_{CLS}^0, z_1^0, \dots, z_K^0] \in \mathbb{R}^{(K+1)\times d},$$

де перший елемент – службовий [CLS].

У шарі $\ell = 1, \dots, L_s$ просторового перетворювача оновлення має вигляд (pre-LN):

$$\tilde{Z}^\ell = Z^{\ell-1} + MSA\left(LN\left(Z^{\ell-1}\right)\right), \quad Z^\ell = \tilde{Z}^\ell + MLP\left(LN\left(\tilde{Z}^\ell\right)\right),$$

де $Z^{\ell-1} \in \mathbb{R}^{(K+1)\times d}$ – послідовність ознакових векторів на вході шару ℓ; перша рівність – увага з попередньою нормалізацією та залишковим додаванням; друга – позиційно-незалежне MLP з нормалізацією та залишковим додаванням».

У кожному шарі просторового трансформеру (із h каналами уваги, L_s шарів, коефіцієнт розширення r) застосовується

$$Attn(Q,K,V) = softmax\left(\frac{(Q\,K^{\mathrm{T}})}{\sqrt{d_h}}\right)V,$$

де $d_h = \frac{d}{h}$, а $Q, K, V \in \mathbb{R}^{(K+1)\times d_h}$ – проєкції Z для одного каналу уваги. Після блоків отримано CLS-вектор кадру c_t

Баланс класів і навчання. Клас-зважена перехресна ентропія або фокальна (зважувальна) втрата (параметр $\gamma \approx 2$); зважене випадкове формування пакетів даних (аналог WeightedRandomSampler); вибір T кадрів за способом часових сегментів (TSN).

Штучні варіації даних: горизонтальне віддзеркалення, помірна зміна яскравості/контрасту.

Оптимізація: оптимізатор AdamW (швидкість навчання $\approx 3 \cdot 10^{-4}$), штраф за ваги 10^{-4} ... 10^{-2}, обрізання градієнта 1.0, дострокове зупинення за Macro-F1.

Функції втрат. Крос-зважена перехресна ентропія:

$$\mathcal{L}_{CE} = -\left(\frac{1}{N}\right)\sum_{i=1}^{N} w_{y_i} \log(p_{i,y_i} + \varepsilon); \quad w_c = \frac{N}{C \cdot N_c}.$$

Фокальна (зважувальна) втрата:

$$\mathcal{L}_{Focal} = -\left(\frac{1}{N}\right)\sum_{i=1}^{N} w_{y_i}(1 - p_{i,y_i})^{\gamma} \log(p_{i,y_i} + \varepsilon),\ \gamma = 2\,,$$

де $p_{i,c} = \mathrm{softmax}(z_i)_c$ $y_i \in \{1, \dots, C\}$; N_c – кількість прикладів класу c у тренувальній вибірці $\varepsilon = 10^{-7}$ – мале додатне число для числової стабільності, додається в $\log(p + \varepsilon)$, щоб уникнути log (0).

Зважене формування пакетів даних. Використовуються ваги прикладів, пропорційні w_{y_i}.

Вибір T кадрів. TSN-схема: кліп ділиться на T рівних **часових сегментів**, з кожного береться один кадр (випадковий або центральний).

Прискорення. Величини оптичного потоку та сформовані адаптивні фрагменти зберігаються у кеші на диску; епохи на CPU пришвидшуються приблизно у 3-5 разів порівняно з обчисленням «на льоту».

Експериментальна частина

Дані та препроцесинг

- CCD1500: 1500 аварійних і 3000 нормальних відеороликів (.mp4).
- Поділ на підмножини: навчальна 80% / валідаційна 20% із збереженням часток класів.
- Масштабування кадрів: 112×112 (варіант — 128×128)

- Відбирання кадрів (рівномірно по всій довжині кліпу). Обирається T =6…8 кадрів.

$$t_i = \left(i * \frac{L-1}{T-1} \right), i = 0..T-1$$

- Оптичний потік (Farnebäck, OpenCV). Параметри:
  $\text{pyr}_{\text{scale}} = 0.5, \text{levels} = 3, \text{winsize} = 15, \text{iterations} = 3, \text{poly}_{\text{n}} = 5, \text{poly}_{\text{sigma}} = 1.2$

Використано модуль вектору потоку між кадрами X_t та X_{t+1}:

$$M_{t(x,y)} = \sqrt{u_{t(x,y)}^2 + v_{t(x,y)}^2},$$

де u, v – компоненти потоку Farnebäck між X_t та X_{t+1}.
Модулі mag_{t} кешуються на диск для пришвидшення навчання.

Метрики для порівняння результатів

Використано Точність (Accuracy), Повноту (Recall), Точність позитиву (Precision), усереднений за класами F1 (Macro-F1).

- Accuracy

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}.$$

- Precision та Recall для кожного класу *c*

$$\text{Prec}_{\text{c}} = \frac{\text{TP}_{\text{c}}}{\text{TP}_{\text{c}} + \text{FP}_{\text{c}}}, \text{Rec}_{\text{c}} = \frac{\text{TP}_{\text{c}}}{\text{TP}_{\text{c}} + \text{FN}_{\text{c}}}.$$

- Macro-F1 (середнє F1 по класах, де *C* — кількість класів)

$$\text{MacroF1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot Prec_c \cdot Rec_c}{Prec_c + Rec_c}.$$

Результати та порівняння з існуючими підходами

У таблиці наведено результати роботи описаного методу та інших вже існуючих методів на датасеті CCD1500. Для порівняння було використано:

- той самий поділ вибірки (80/20, зі збереженням пропорцій класів), та сама роздільна здатність 112×112 та однакові перетворення даних;
- баланс класів: однаково – крос-ентропія з вагами класів або фокусована функція втрат плюс вибірка зі зваженою ймовірністю;
- раннє зупинення за Macro-F1, однакове обмеження кількості епох, експеримент повторюється N разів із різними початковими ініціалізаціями;
- для огляду результатів використано: кількість фрагментів на один кадр, кількість параметрів (млн), ознакові одиниці на кадр (для трансформерних варіантів), частка правильних відповідей (Accuracy), макро-F1, Precision, Recall кадрів за секунду (ЦП/ГП).

У таблиці порівняно Adaptive-Sparse-ViT (описаний у роботі, керований рухом), Uniform-patch ViT (без урахування руху) та TSN (ResNet-18). Умови однакові: спліт 80/20 зі збереженням пропорцій класів, роздільна здатність 112×112, однакові перетворення даних, фіксована схема TSN з однаковим *T* сегментів, однакова політика навчання (AdamW, рання зупинка за Macro-F1), фіксовані значення генератора випадкових чисел. Для описаного у роботі методу середня кількість фрагментів на кадр ≈96 ± 18 при верхній межі $K_{\max} = 120$ (проти 196 у рівномірної сітки); кількість параметрів у обох ViT ≈1.2 млн, у TSN – ≈11.7 млн. На етапі валідації CCD1500 запропонований підхід досягає accuracy = 0.864 і macro-F1 = 0.851, перевершуючи Uniform-patch ViT (0.83 /0.80) та TSN (0.82 / 0.78) за збереження нижчої обчислювальної вартості. Необхідно зазначити, що значення наведені як середнє ± стандартне відхилення за кількома запусками.

Таблиця. **Результати роботи описаного методу та інших вже існуючих методів на датасеті CCD1500**

Метод	**Фрагментів /кадр (b=8, 112×112)**	**Кількість параметрів, млн**	**Accuracy**	**Macro-F1**	**Precision**	**Recall**	**Кадрів/с (ЦП/ГП)**
Adaptive-Sparse-ViT (описаний у роботі, керований рухом)	~80–120 (≤0.6·N_base)*	~1.2–1.5	**0.864**	**0.851**		**0.860**	≈12–18 / ≈220–280
Uniform-patch ViT (без урахування руху)	196 (=N_base)	**~1.0–1.2**	0.83	0.80	0.80	0.78	≈8–12 / ≈160–220
TSN (ResNet-18)	–	~11.0–12.0	0.82	0.78	0.81	0.79	≈8–12 / ≈160–220

Висновки та подальші кроки. У роботі запропоновано поєднання адаптивного відбору фрагментів (рух → масштаб фрагмента) з двоступеневим просторово-часовим перетворювачем, в результаті чого вдалося знизити обчислювальні витрати під час навчання та при застосуванні навченої моделі.

У результаті навчання моделі на наборі даних CCD1500 отримано accuracy = 0.864 і macro-F1 = 0.851 з добрим recall для класу відео Crash (аварійні випадки) за меншої кількості ознакових векторів на кадр (порівняно з рівномірною сіткою). В подальших дослідженнях плануються наступні кроки**:**

- Використання динамічних порогів $\frac{q1}{q2}$ та/або адаптивна верхня межа кількості токенів K_{max} залежно від сцени.
- Застосування стійкіших ознак руху: усереднення кількох оцінок потоку, стабілізація кадру, приглушення тремтіння камери.
- Багатомасштабний контекст: поєднання фрагментів розмірів $b, 2b, 4b$ із взаємодією між масштабами.
- Довший часовий контекст T із пам’яттю (кешування ключів/значень) або двонаправленим агрегуванням.
- Зменшення хибних спрацьовувань (FP): калібрування ймовірностей (temperature scaling), класозалежні пороги.
- Запуск на пристроях: перевести модель у 8-бітний варіант, збереження у формат ONNX або під TensorRT для швидкого виконання, вимірювання час відгуку та використання пам’яті на цільовому пристрої.

СПИСОК ЛІТЕРАТУРИ

1. Dosovitskiy A. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. *ICLR*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929.

2. Bertasius G., Wang H., Torresani L. “Is Space-Time Attention All You Need for Video Understanding?” *ICML*. 2021. DOI: https://doi.org/10.48550/arXiv.2102.05095.

3. Arnab A., Dehghani M., Heigold G., Sun C., Lučić M., Schmid C. “ViViT: A Video Vision Transformer”. *ICCV*. 2021. DOI: https://doi.org/10.48550/arXiv.2103.

4. Liu Z., Ning J., Cao Y., Wei Y., Zhang Z., Lin S., Hu H. “Video Swin Transformer”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2106.13230.

5. Fan H., Xiong B., Mangalam K., Li Y., Yan Z., Malik J., Feichtenhofer C. “Multiscale Vision Transformers”. *arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2104.11227.

6. Ilg E., Mayer N., Saikia T., Keuper M., Dosovitskiy A., Brox T. “FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks”. *IEEE*. Honolulu, HI, USA. 2017. DOI: https://doi.org/10.1109/CVPR.2017.179.

DOI: https://doi.org/10.15276/ict.02.2025.20
UDC: 004.93

Adaptive video streams fragment analysis for traffic accident classification using a sparse video transformer

Tetiana V. Normatova [1)]
PhD Student of the Department of Informatics
ORCID: https://orcid.org/0009-0004-3503-6350, tetiana.normatova@nure.ua
Sergii V. Mashtalir [1)]
Doctor of Engineering Sciences, Professor of the Department of Informatics
ORCID: https://orcid.org/0000-0002-0917-6622; Scopus Author ID: 36183980100
[1)] Kharkiv national university of radio electronics, 14, Nauky Ave. Kharkiv, 61166, Ukraine

ABSTRACT

This paper proposes a simple in software implementation and effective approach to classifying short video fragments into emergency and normal scenes. For this work, a uniform sampling of frames from the entire clip (6–8 frames) is used in order not to lose key events even in long videos. Next, adaptive motion-driven frame "fragmentation" is applied: quantile thresholds are calculated using the Farnebäck optical flow map and the fragment size (8/16/32 pixels) is selected for each cell of the base grid. In areas with pronounced motion, small fragments are selected (higher detail), in static ones - larger ones (fewer calculations). The selected fragments do not overlap, are scaled to the base size and converted into feature vectors. The architecture is built on a two-stage principle. In the first step, the spatial attention block works within one frame — only on selected fragments, which significantly reduces the number of feature units. In the second step, the temporal block processes the sequence of frames through their short summary representations (service classification markers, hereinafter – CLS), aggregating dynamics in time. Such a factorization "space → time" reduces the computational cost and memory without losing informativeness in moving regions. To combat class imbalance, a weighted loss function (or "loss with a focus on heavy examples") and random sampling with weights are used during training. Optical flow maps and lists of selected fragments are previously stored on disk, which accelerates epochs on the processor without special graphics equipment. The evaluation is performed on CCD1500 (1500 emergency and 3000 normal videos) with a standard 80/20 division while preserving class fractions. The accuracy is 0.864 and the macro-F1 is 0.851; according to the preliminary comparison, the proposed approach outperforms the basic uniform frame splitting and classical schemes with simple time sampling. The main value of the approach is the combination of "motion-driven" feature unit reduction and two-stage processing, which makes the model suitable for realistic time and resource (CPU) constraints, while maintaining high sensitivity to short and local emergency events. The method is easy to scale and combine with pre-training based on masked video reconstructions. The condition fixation, open settings, and steps for full reproducibility are also described.

Keywords: Video classification; neural networks; convolutional neural networks; object classification; video stream analysis; data classification; image fragment processing