https://doi.org/10.31891/2307-5732-2025-353-4
УДК 637.5.02

ЯРЕМЧЕНКО ОЛЕКСАНДР
Національний університет «Львівська політехніка»
https://orcid.org/0009-0001-2002-2704
e-mail: Oleksandr.D.Yaremchenko@lpnu.ua
ПУКАЧ ПЕТРО
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-0359-5025
e-mail: Petro.Y.Pukach@lpnu.ua

РОЗПІЗНАВАННЯ МІКРОВИРАЗІВ ЗА ДОПОМОГОЮ АРХІТЕКТУРИ ТРАНСФОРМЕРА

Міміка тісно пов'язана з рухами та скороченнями мімічних м'язів, де чіткі м'язові активації відображають різні емоційні стани. У випадку мікровиразів — коротких, мимовільних виразів обличчя — ці рухи м'язів надзвичайно тонкі й швидкоплинні, часто тривають менше півсекунди. Ця тонкість становить серйозну проблему для сучасних алгоритмів розпізнавання емоцій на обличчі, багато з яких розроблені для більш відвертих і тривалих виразів. Як наслідок, існуючим моделям часто важко розпізнати мікровирази через низьку інтенсивність і коротку тривалість мімічних сигналів. Багато найсучасніших підходів використовують механізми самоуважності для моделювання зв'язків між токенами в часовій послідовності. Однак у цих моделях зазвичай не враховуються внутрішні просторові відносини між орієнтирами на обличчі, які є важливими для розуміння дрібнозернистих м'язових рухів, залучених до мікровиразів. Відсутність просторового усвідомлення може призвести до неоптимальної продуктивності, особливо при спробі виявити мінімальну та локалізовану м'язову активність. Щоб вирішити цю проблему, ми пропонуємо нову мережу ієрархічних трансформаторів (HTNet), спеціально розроблену для покращення розпізнавання мікровиразів шляхом більш ефективного вивчення локалізованої динаміки м'язів обличчя. HTNet складається з двох основних компонентів: трансформаторного рівня, який фіксує локальні часові особливості, та агрегаційного рівня, який витягує як локальні, так і глобальні семантичні представлення активності обличчя. Модель розділяє обличчя на чотири ключові області: ліва губа, права губа, ліве око та праве око. Кожна область обробляється трансформаторним шаром незалежно, використовуючи локалізовану увагу на собі, щоб зосередитися на незначних рухах м'язів. Потім рівень агрегації вивчає міжрегіональні взаємодії, особливо між областями очей і губ. Наші експерименти, проведені на чотирьох широко використовуваних наборах даних мікровиразів, демонструють, що HTNet значно перевершує існуючі методи, встановлюючи новий стандарт точності та надійності завдань розпізнавання мікровиразів.

Ключові слова: ієрархічний трансформатор, розпізнавання мікровиразів, глибоке навчання, рух м'язів обличчя, локальна самоувага.

YAREMCHENKO OLEKSANDR, PUKACH PETRO
Lviv Polytechnic National University

MICRO-EXPRESSION RECOGNITION USING TRANSFORMER ARCHITECTURES

Facial expressions are closely linked to the movements and contractions of facial muscles, where distinct muscle activations reflect different emotional states. In the case of micro-expressions—brief, involuntary facial expressions—these muscle movements are extremely subtle and fleeting, often lasting less than half a second. This subtlety poses a significant challenge for current facial emotion recognition algorithms, many of which are designed for more overt and prolonged expressions. As a result, existing models often struggle with micro-expression recognition due to the low intensity and short duration of the facial cues. Many state-of-the-art approaches employ self-attention mechanisms to model relationships between tokens in a temporal sequence. However, these models typically overlook the intrinsic spatial relationships among facial landmarks, which are essential for understanding the fine-grained muscle movements involved in micro-expressions. The lack of spatial awareness can lead to sub-optimal performance, especially when trying to detect minimal and localized muscle activity. To address this, we propose a novel Hierarchical Transformer Network (HTNet), specifically designed to enhance the recognition of micro-expressions by learning the localized facial muscle dynamics more effectively. HTNet consists of two core components: a transformer layer that captures local temporal features and an aggregation layer that extracts both local and global semantic representations of facial activity. The model partitions the face into four key regions: left lip, right lip, left eye, and right eye. Each region is processed independently by the transformer layer using localized self-attention to focus on minor muscle movements. The aggregation layer then learns the inter-regional interactions, especially between the eye and lip areas. Our experiments, conducted on four widely used micro-expression datasets, demonstrate that HTNet significantly outperforms existing methods, establishing a new benchmark for accuracy and robustness in micro-expression recognition tasks.

Keywords: Hierarchical transformer, Micro-expression recognition, Deep learning, Facial muscle movement, Local self-attention.

Стаття надійшла до редакції / Received 07.04.2025
Прийнята до друку / Accepted 22.04.2025

Вступ

Мікровираз - це тонкий рух м'язів, що триває лише 0,04 секунди. В останні роки були проведені різні типи досліджень щодо використання методів комп'ютерного зору для аналізу мікровиразів [1]. Однак точність, з якою розпізнаються мікровирази, ще потребує вдосконалення. Хоча набір даних для мікроекспресії збирається в лабораторних умовах і лабораторне середовище добре контролюється, найкращий результат для мікроекспресії є незадовільним [2]. Використання методів комп'ютерного зору все ще є складним завданням. Проте звичайне розпізнавання макровиразів досягло високої

точності розпізнавання (більше 95%)[3]. Причина в тому, що мікровирази часто супроводжуються тонкими рухами м'язів, що ускладнює виявлення мікровиразів людьми та комп'ютерами. Щоб виділити риси обличчя, деякі дослідники в галузі розпізнавання мікровиразів запропонували використовувати локальний бінарний шаблон (LBP). Методи LBP мають хорошу здатність розрізняти за допомогою методу виділення ознак на основі текстури та низьку обчислювальну складність. Крім того, інші дослідники використовують характеристики оптичного потоку як вхідні дані для оцінки руху м'яза [4]. Оптичний потік визначається різницею яскравості між кадрами, які використовуються для оцінки тонких рухів обличчя. Методи, засновані на оптичному потокі, включають Bi-WOOF[5], MDMO[6], FHOFO [7], вагу оптичної деформації та характеристику оптичної деформації[8]. VGG16[9], GoogleNet[10], AlexNet[11] і OFF-Apex[2] отримують оптичний потік TV-L1 як вхідні дані з вибраних кадрів вершини та початку послідовності зображень. Піковий кадр фіксує найбільш помітну інформацію про мікровираз, і його можна використовувати для точного розпізнавання виразу. Двома основними етапами розпізнавання мікровиразів є класифікація рис обличчя та витяг релевантної інформації про обличчя із зображень. Традиційна класифікація здебільшого спирається на ознаки, які були створені штучно. Методи глибокого навчання використовують зображення обличчя для автоматичного визначення рис. Щоб зрозуміти тонкий рух обличчя та отримати рішення на основі обмежених навчальних даних, Лю та ін. [4] пропонують використовувати методи адаптації домену для передачі знань про макровирази для розпізнавання мікровиразів. Під час процесу мережа могла збільшувати мікровираз і додавати більше згенерованих зображень до навчального набору даних, а саме збільшення та зменшення виразу (EMR). Продуктивність цієї системи конкурентоспроможна на Micro-Expression Grand Challenge. Ліонг та ін. запропонуйте використовувати STSTNet, який є трьома неглибокими потоками CNN, щоб розпізнавати вираз обличчя. Поля оптичної деформації, вертикального оптичного потоку та горизонтального оптичного потоку подаються в ці три мережі дрібних потоків [12]. З метою розпізнавання мікровиразів Xia et al. пропонують прийняти рекурентну згортку мережу, яка має менший дизайн і поєднує в собі сильні сторони як RNN, так і CNN[13]. Рух для мікроекспресії часто невеликий і його важко виявити. Це ускладнює ідентифікацію конкретних м'язів обличчя, які беруть участь у виразі обличчя, і відстеження руху пікселя з часом. Щоб вирішити ці проблеми, Zhou et al. [14] пропонують використовувати мережу уточнення ознак, мережу навчання особливостей специфічного виразу та методи злиття для розпізнавання виразів. Використовуючи самоувагу на глобальних характеристиках зображення, їхня техніка може виділити відмінні характеристики для розпізнавання мікровиразів. Однак глобальне самозвернення уваги між елементами зображення є дорогим з обчислювальної точки зору. Зменшення розміру самоуважності є більш обчислювально ефективним під час навчання моделі [15] [16]. Тому багато останніх зусиль намагаються знизити вартість обчислень шляхом застосування самоконтролю до локальних блоків і грубого самоконтролю до глобальних блоків. Локальна та глобальна увага може охопити візуальну залежність як зблизька, так і з віддаленої сторони, що може ефективно підвищити ефективність узагальнення та стійкість моделі. Наша мета полягає в тому, щоб зосередити увагу на локальних блоках на кожному рівні ієрархії, охопити дрібнозернисті функції та використовувати блоки агрегації для включення як локальних дрібнозернистих, так і глобальних грубозернистих взаємодій у різних масштабах зображення. У цьому новому механізмі кожен піксель у однорівневих блоках має дрібну деталізацію, а пікселі в блоках верхнього рівня мають грубу деталізацію. Використовуючи цей механізм, наша модель може ефективно записувати як короткострокові, так і дальні візуальні залежності.

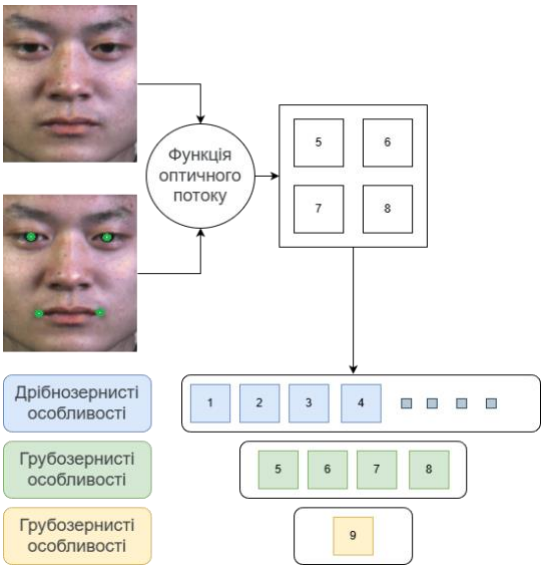

Рис. 1: Запропонована парадигма HTNet. Запропонована структура зосереджена на чотирьох областях обличчя замість цілих областей обличчя — області лівого ока, області правого ока, області лівих губ і області правих губ, що може усунути звуковий фоновий шум, який уловлює лабораторна камера через потенційне мерехтіння світла

Зокрема, у цій роботі ми представляємо новий метод самоконтролю шарів трансформера для захоплення локальних і глобальних взаємодій в ієрархічній структурі. Низькорівнева самоувага в трансформерних шарах вловлює дрібні особливості в локальних регіонах. Високорівнева самоувага в шарах трансформера вловлює грубозернисті особливості в глобальних регіонах. Блок агрегації пропонується для створення взаємодії між різними блоками на одному рівні. Загальна парадигма HTNet описана на рис.1. Наш внесок можна підсумувати таким чином:

a) Ми представляємо новий механізм самоконтролю для трансформерних рівнів для захоплення як локальних, так і глобальних взаємодій в ієрархічній структурі, що дозволяє розпізнавати мікровирази в зображеннях за допомогою запропонованої функції агрегації блоків. Низькорівнева та високорівнева самоувага може вловлювати дрібнозернисті та грубозернисті риси в локальних та глобальних регіонах.

b) Наша мережа зосереджена на чотирьох зонах обличчя замість цілих областей обличчя — області лівого ока, області лівих губ, області правого ока та області правої губи, — що може пом'якшити вплив помітного фонового шуму, який уловлює лабораторна камера через можливе мерехтіння світла. Трансформуючий шар використовується для зосередження на моделюванні невеликих тонких м'язових рухів із локальною самоувагою в кожній області. Крім того, рівень агрегації вивчає взаємодію між областями очей і губами. Ми перевіряємо, як різні розміри блоків впливають на точність розпізнавання мікровиразів.

c) Експерименти на чотирьох доступних наборах даних демонструють, що наш метод помітно перевершує попередні підходи.

Аналіз досліджень та публікацій

Традиційні методи

Для розпізнавання виразів у звичайних техніках часто використовуються характеристики, засновані на зовнішності. Найпоширенішим шаблоном є локальний бінарний шаблон із трьох ортогональних площин (LBP-TOP)[17]. В інших роботах LBP-TOP LBP-TOP переноситься в тензорно-незалежний простір RGB, який є більш надійним [18]. Менша обчислювальна складність досягається використанням LBP-SIP для ефективної мінімізації надмірності в шаблонах LBP-TOP і забезпечення легкого представлення [19]. LSDF [20] використовує 16 регіонів інтересу (ROI), які система кодування дій обличчя може виділити для розпізнавання мікровиразів. 16 регіонів інтересу (ROI) представляють 16 окремих областей руху м'язів, виключаючи невідповідні області. STCLQP [21] містить додаткові функції, включаючи орієнтацію, атрибути форми та величину. Включаючи компоненти знака, величини та орієнтації, STCLQP розширився від LSDF і зменшив зони інтересу з 16 до 3. Кодову книгу створено для вилучення дискримінаційних і помітних ознак із кодових книг для різних емоцій обличчя. У традиційних методах геометричні об'єкти виділяються за допомогою орієнтирів обличчя або оптичного потоку. Геометричні ознаки можуть ідентифікувати деформації руху. Лі та ін. [22] досліджували техніку глибокого навчання для локалізації орієнтирів обличчя та розділили область обличчя на зони інтересу. Мікровираз обличчя створюється скороченнями м'язів обличчя; тому оцінка орієнтації м'язових скорочень обличчя важлива для ідентифікації емоцій. Використовуючи цю техніку, вони класифікували обличчя на різні зони, що викликають занепокоєння, які корелюють з різними моделями дій м'язів. Крім того, оптичний потік може адекватно відображати активність області обличчя. Лю та ін. [6] пропонують використовувати потужну мережу виділення ознак MDMO для ідентифікації мікровиразів. Вони обчислюють оптичний потік для кожного зображення у відеоряді та поділяють зони обличчя на численні чіткі та інтригуючі частини. Після цього розраховуються середні характеристики оптичного потоку для кожного зображення; класифікатор SVM буде використовуватися для цих середніх оптичних характеристик потоку для ідентифікації емоцій. Їхній метод може враховувати як географічне положення, так і регіональну статистичну інформацію про рух. Їх метод простий і ефективний. Замість використання гістограми LBP слід використовувати пікову рамку для передачі вирішального вираження. Liong та ін. [5] запропонували зважити гістограму LBP та усереднити значення характеристик оптичного потоку для розпізнавання емоцій на обличчі, а саме Bi-WOOF.

Методи глибокого навчання

Останніми роками було впроваджено численні методи глибокого навчання для вилучення характеристик обличчя для ідентифікації виразу [23, 24]. Хіа та ін. [25] запропонували використовувати рекурентні згорткові мережі для з'єднання інформації про положення обличчя та запису скорочення м'язів обличчя між різними м'язовими регіонами. Під час процесу дрібні варіації мікровиразів будуть збільшені. Ця модель включає кілька повторюваних згорткових шарів і шар класифікації, які будуть використовуватися для захоплення візуальних характеристик для ідентифікації емоцій обличчя. Щоб дізнатися про зв'язки між вузлами та краями обличчя, Lei et al. [23] запропонував використовувати Transformer як кодер. Графік обличчя створюється за допомогою ребра та вузлів рис обличчя. Більшість попередніх методів використовують елементи ручної роботи для відображення тонких скорочень обличчя. Однак ручна робота вимагає більших обчислювальних витрат, тому Gan et al. [2] запропонував автоматичну локалізацію вершинного кадру та використання оптичного потоку як вхідних даних для

OFF-ApexNet. Мережа може отримувати нові дескриптори функцій з оптичного потоку за допомогою CNN. Більшість звичайних методів часто віддають перевагу використанню глибших мереж і складних об'єктивних функцій для вилучення розширених візуальних функцій. Однак глибша нейронна мережа часто не може підвищити ефективність моделі через проблеми з переобладнанням. Отже, деякі дослідники розробили неглибоку CNN, щоб отримати високорівневі візуальні атрибути з трьох компонентів оцінки руху – горизонтального, вертикального полів оптичного потоку та оптичного напруження для визначення емоційних станів [12]. Щоб вирішити проблему зсуву домену композитної бази даних, Xia et al. [13] розробили модель RCN, щоб дослідити, як менша модель впливає на виявлення мікровиразів. Вони розробили три модулі без параметрів у RCN, такі як блок уваги, швидке підключення та широке розширення. Використовуючи ці три модулі без параметрів, параметри, які можна вивчати, не збільшуються. Інші дослідники помітили, що ієрархічна структура є важливою [26].

Трансформер у комп'ютерному зорі

Стандартна увага, вперше представлена в Transformer, — це масштабована увага за скалярним добутком, яка стала основним методом у завданнях НЛП. Vision Transformer (ViT) [27] представляє Transformer від завдань НЛП до завдань комп'ютерного зору. Для класифікації зображень Vision Transformer (ViT) застосовує архітектуру Transformer до блоків зображень, які не перекриваються. Однак Vision Transformer (ViT) потребує величезних навчальних наборів даних, щоб покращити продуктивність своєї мережі. Vision Transformer (ViT) тепер може ефективно працювати з меншими навчальними наборами даних завдяки деяким новим методологіям навчання, представленим DeiT. Однак Vision Transformer (ViT) не підходить як магістральна мережа для щільних завдань підсистеми бачення, оскільки зі збільшенням розміру зображення обчислювальна складність квадратично зростає. Щоб застосувати трансформерну архітектуру у високій роздільній здатності пікселів у зображеннях, Лю та ін. [28] запропонував ієрархічний трансформер, відомий як Swin Transformer, який отримується шляхом зсуву вікон. Обмежуючи обчислення власної уваги до локального вікна, одночасно дозволяючи міжвіконний зв'язок у різних блоках зображення, зміщені вікна можуть підвищити ефективність моделі. Інша пов'язана робота Transformer [29], [30] досліджує прості магістральні мережі без згорток для завдань глибокого прогнозування.

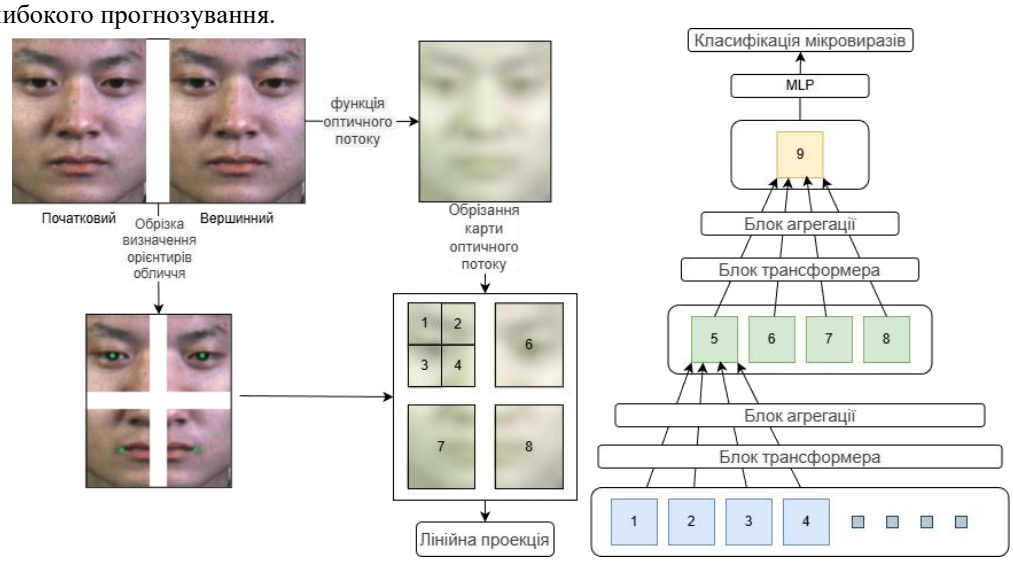

Рис. 2: HTNet: Загальна архітектура ієрархічної трансформерної мережі для розпізнавання мікровиразів. Низькорівнева самоувага в трансформерних шарах вловлює дрібні особливості в локальних регіонах. Високорівнева самоувага в шарах трансформера вловлює грубозернисті особливості в глобальних регіонах. Блок агрегації пропонується для створення взаємодії між різними блоками на одному рівні

Запропонований метод

Згідно з рисунком 2, запропонована HTNet включає трансформерні рівні та агрегацію блоків на кожному рівні ієрархічної структури. Трансформуючий шар проводить самоконтроль на кожному блоці зображення незалежно. Дрібнозернисті характеристики вловлюються на низькорівневій мережі за допомогою функції самоконтролю на рівні трасформера. Після цього агрегація блоків об'єднає малі блоки зображень у більший блок зображення. Створюються взаємодії між різними блоками на одному рівні, а детальні характеристики фіксуються після кожного агрегування блоків. Усі блоки на одному рівні мають однаковий список параметрів. Нарешті, блок MLP у нашій моделі застосовується до остаточних карт ознак для класифікації мікровиразів.

Вилучення карти оптичного потоку

Оптичні потоки генеруються з початкового та пікового кадрів і зазвичай використовуються для опису переміщень руху лицьових областей. Вони вже показали багатообіцяючі результати в наборах даних для розпізнавання мікровиразів [2, 12]. Індекс початку та вершини кадру слід визначити перед отриманням карти оптичних ознак. З відеопослідовностей потрібно визначити лише індекс верхнього

кадру, оскільки індекс початкового кадру вже присутній у наборах даних мікровиразів. У нашій статті ми використовуємо метод D&C-RoIs [31] для виявлення пікового кадрового індексу як попередні дослідження мікроекспресії [2]. Підхід D&C-RoIs визначає зв'язок між кадром початку та наступними кадрами:

$$d = \frac{\sum_{i=1}^{B} h_{1i} \times h_{h2i}}{\sqrt{\sum_{i=1}^{B} h_{1i}^2 \times \sum_{i=2}^{B} h_{2i}^2}}, \quad (1)$$

де B — кількість бінів на гістограмах, h_1 — перший кадр, а h_2 — інші кадри, крім першого. Буде вибрано найвищу швидкість різниці в ROI, що представляє рамку верхівки з максимальними змінами м'язів обличчя. Потім ми отримуємо карти характеристик оптичного потоку, використовуючи кадри початку та піку. Карту характеристик оптичного потоку можна сформулювати наступним чином:

$$V = (u(x,y), v(x,y)) | x = 1,2,\dots,X, y = 1,2,\dots,Y, \quad (2)$$

де X і Y представляють ширину кадру W і висоту H відповідно, $u(x,y)$ і $v(x,y)$ — горизонтальний і вертикальний компоненти карти функцій оптичного потоку V, V = [Vx, Vy], де V ∈ RW×H×2 Ми використовуємо похідні першого порядку від поля оптичного потоку, щоб обчислити варіацію полів оптичного потоку або оптичну деформацію, яка може приблизно відображати ступінь зміщення обличчя:

$$V_z = \sqrt{\frac{\partial V_x^2}{\partial x} + \frac{\partial V_y^2}{\partial y} + \frac{1}{2}\left(\frac{\partial V_x^2}{\partial y} + \frac{\partial V_y^2}{\partial x}\right)}, \quad (3)$$

де $\frac{\partial V_x^2}{\partial x}$, $\frac{\partial V_y^2}{\partial y}$, $\frac{\partial V_x^2}{\partial y}$ та $\frac{\partial V_y^2}{\partial x}$ часткові похідні першого порядку V. Нарешті, тривимірні карти характеристик оптичного потоку формуються та представляють як Vm = [Vx, Vy, Vz] і $V_m \in R^{W \times H \times 3}$.

Наша мережа зосереджується на області лівого ока, області лівого ока, області правого ока та області правої губи замість усієї області обличчя, що може усунути звуковий фоновий шум, записаний камерою лабораторії через потенційне мерехтіння світла. Ми використовуємо MTCNN [32], щоб отримати координати орієнтирів обличчя із зображень вершини. Чотири карти характеристик оптичного потоку на обличчі вирізані з усіх карт V_m характеристик оптичного потоку. Чотири карти характеристик оптичного потоку обличчя: карта характеристик оптичного потоку лівого ока, карта візуальних характеристик лівої губи, карта функцій оптичного потоку правого ока та карта функцій правої губи. Розмір цих чотирьох карт характеристик оптичного потоку становить $\frac{W}{2} \times \frac{H}{2} \times 3$, що становить половину розміру всього зображення оптичного потоку. Після цього комбіновані чотири функції оптичного потоку подаються в нашу HTNet. Процес зображено на рис. 2.

Трансформерний шар

Кожен блок зображення має розмір P × P, коли використовується вхідне зображення оптичного потоку розміром H × W × 3. Розмір ознаки кожного патча становить P × P × 3 після лінійної проекції та розділення на зображеннях оптичного потоку, після зведення вхідними даними для нашої моделі є $X \in R^{b \times Hn \times n \times d}$, де H_n – кількість блоків на кожному рівні, b – розмір партії, n – довжина послідовності. $H_n \times n = H \times W/P^2$. У кожному блоці буде застосовано кілька трансформерних шарів. Ієрархія визначає кількість шарів трансформера. Рівень трансформера складається з рівня нормалізації (LN), рівня самоконтролю з кількома головками (MSA) і повнопідключеної мережі прямого зв'язку (FFN). Через включення вектора позиційного вбудовування, який можна навчити, у всі вектори послідовності Rd, просторова інформація буде закодована:

$$\begin{aligned} Y_l'^i &= Y_{l-1}^i + MSA(LBN(Y_{l-1}^i)) \\ Y_l^{i+1} &= Y_l'^i + FFN(LN(Y_l'^i)) \\ Y_l'^{i+1} &= Y_l^{i+1} + MSA(LN(Y_l^{i+1})) \\ Y_l^{i+2} &= Y_l'^{i+1} + FFN(LN(Y_l'^{i+1})), \end{aligned} \quad (4)$$

де l = 1, 2, ..., L, l – індекс l − го блоку в кожному рівні ієрархії i, а L – загальна кількість блоків у кожному рівні ієрархії. FFN включає два рівні: $\max(0, xW_1 + b)W_2 + b$. Для кожного блоку і на тому самому рівні буде застосовано багатоголове самоувага. У частині самоконтролю входи $X \in R^{n \times d}$ перетворюються на три частини: запити Q, ключі K і значення V, де n — довжина послідовності, а d — розмірність входів. Масштабований скалярний добуток буде застосовано до Q, K, V:

$$MSA(Q,K,V) = soft\,max\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (5)$$

Нормалізація рівня є важливою частиною трансформера, яка може зробити навчання мережі стабільним і швидкою конвергенцією. LN буде застосовано в кожному блоці наступним чином:

$$LN(x) = \frac{x - \mu}{\delta} o\lambda + \beta, \quad (6)$$

де μ — середнє значення ознак, а δ — стандартне відхилення ознаки, o — поелементна точка, а λ і β — параметри, які можна дізнатися. Після рівня трансформера агрегація блоків буде застосована до виходу рівня трансформера. Кожні чотири маленькі блоки будуть об'єднані в один більший блок.

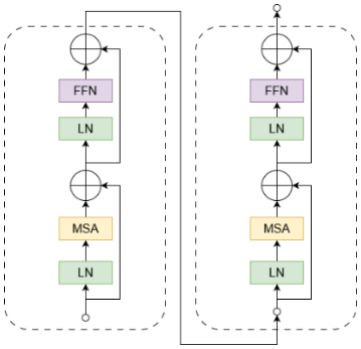

Рис. 3: Кілька трансформерних шарів будуть застосовані в кожному блоці паралельно. Ієрархія визначає кількість шарів трансформера. Рівні трансформера складаються з нормалізації рівня (LN), рівня самоконтролю з кількома головками (MSA) і повнопідключеної мережі прямого зв'язку (FFN). Просторова інформація буде закодована шляхом додавання вектора позиційного вбудовування, який можна навчити, до всіх векторів послідовності в R^d

Блок агрегації

Функція агрегування блоків у нашій HTNet подібна до кількох схем Pyramid. Однак більшість попередніх методів використовують глобальну увагу до зображень. На відміну від попередніх методів піраміди, наша модель використовує локальну увагу до кожного блоку зображення, що може ефективно підвищити ефективність моделі, оскільки зони руху м'язів є локальними особливостями, які зосереджені на конкретних областях обличчя. Більшість ділянок обличчя не мають значення для визначення станів мікровиразу. Окрім уваги на місці, важлива також локальна міжблокова комунікація. Swin-transformer [28] представляє підхід до розділення вікон із зсувом і маскований MSA для створення зв'язків між сусідніми блоками та ефективний у семантичній сегментації, класифікації зображень і виявленні об'єктів. Однак підхід із зміщеним віконним розділенням все ще може ускладнити увагу до себе, і його непросто реалізувати на практиці.

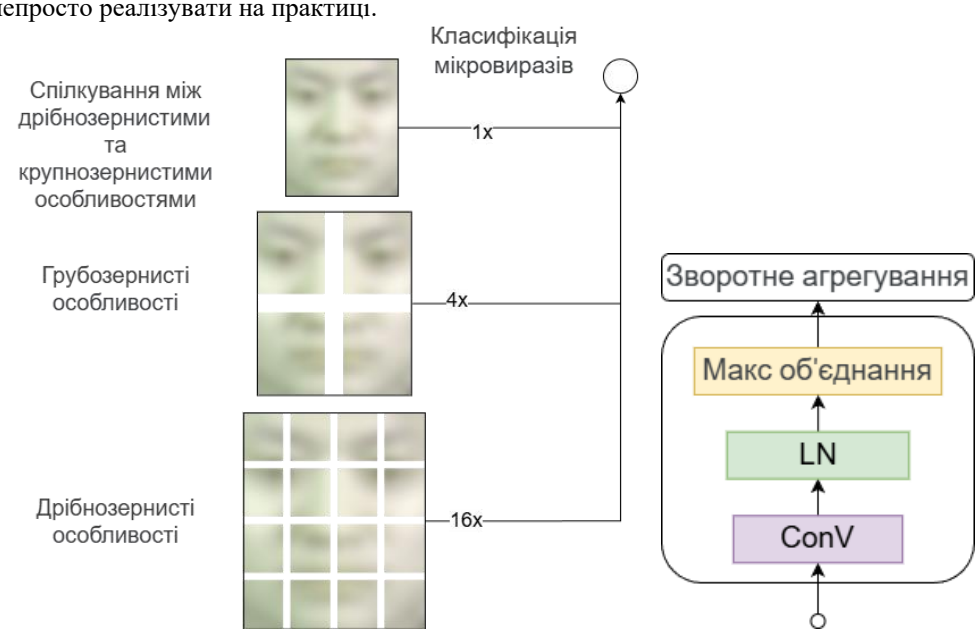

Рис. 4: Агрегація блоків включає згортковий рівень 3 × 3, а потім LN і максимальне об'єднання 3 × 3. У нижній частині нашої моделі карта лицевого оптичного потоку включає 16 лицевих блоків, які є картами ознак 4 × 4. Використовуючи згортковий шар 3 × 3 на цій карті функцій, часткові карти характеристик у межах чотирьох областей обличчя будуть об'єднані, а розмір блоку стане 2 × 2 блоки, що відповідає чотирьом зонам обличчя: області лівого ока, області лівого ока, області правого ока та області правої губи

У нашій моделі HTNet кожен блок незалежно обробляє карту оптичного потоку, а зв'язок між блоками відбувається під час агрегації блоків. Агрегація блоків виконується на чотирьох сусідніх блоках, щоб у цій операції можна було обмінюватися сусідніми й глобальними функціями. Агрегація блоків низького рівня обмінюватиметься інформацією в межах чотирьох областей обличчя замість цілої області обличчя: області лівого ока, області правого ока, області лівої губи та області правої губи. Дрібнозернисті функції будуть витягнуті за допомогою низькорівневої агрегації. У той час як агрегація блоків високого рівня обмінюватиметься глобальною інформацією між чотирма областями обличчя, грубозернисті риси будуть вилучені. Процес можна підсумувати на рис. 4. В ієрархії l зображення меншого оптичного потоку є $X_l \in R^{b\times h\times \omega\times d^l}$. Після агрегації блоків розмір зображення оптичного потоку стає $X_{l+1} \in R^{b\times 2h\times 2\omega\times d^{l+1}}$. Нарешті, усі чотири області обличчя об'єднані в одну карту рис обличчя, тоді як $X \in R^{b\times H\times W\times D}$. У цьому процесі функції добре зберігаються шляхом налаштування dl+1 > dl.

Наша блокова агрегація HTNet включає згортковий рівень 3 × 3, за яким слідують LN і максимальне об'єднання 3 × 3. У нижній частині нашої моделі карта лицевого оптичного потоку включає 16 лицевих блоків, які є картами ознак 4 × 4. Використовуючи згортковий шар 3 × 3 на цій карті функцій, часткові карти функцій у межах чотирьох областей обличчя будуть об'єднані, а розмір блоку стане 2 × 2 блоки, які відповідають чотирьом областям обличчя: області лівого ока, області лівих губ, області правого ока та області правої губи. Знову використовуючи згортковий шар 3 × 3 на цій карті функцій, відбудеться обмін інформацією між чотирма областями обличчя. Розмір блоку стане 1 × 1 блок, і буде видобуто повні карти оптичних функцій. Нарешті, витягнуті повні карти функцій будуть подані на рівень MLP для класифікації мікровиразів.

Функція втрат

У цій статті ми використовуємо функцію втрат крос-ентропії для навчання нашої моделі. Наведену нижче формулу можна використовувати для обчислення крос-ентропійних втрат L:

$$L = \sum_i (-\omega_i \log(p_t^i)) \qquad (7)$$

$$p_t = p_c^y (1 - p_c)^{1-y},$$

де ω_i — вага зразків у наборі даних, y — мітка, а $y_i \in 0, 1$.

Досліди

Набори даних

Експерименти реалізовані в базах даних SAMM[33], SMIC[34], CASME II[35] і CASME III [36]. SAMM, SMIC і CASME II об'єднані в один складений набір даних, і однакові мітки в цих трьох наборах даних застосовуються для завдань мікровиразу. У цих наборах даних категорія «позитивних» емоцій включає клас емоцій «щастя», а категорія «негативних» емоцій включає класи емоцій «сум», «огида», «презирство», «страх» і «гнів», тоді як категорія емоцій «здивування» включає лише клас «сюрприз»

SAMM [33]: набір даних SAMM включає 28 учасників, 133 мікровирази та 147 довгих відео з 343 макровиразами. Набір даних містить комплексні одиниці дії, які закодовані. Початок, зміщення та індекс вершини надаються в SAMM. Початкова роздільна здатність зразків становить 2040 на 1088 пікселів. Частота кадрів становить 200. Ми обрізали зображення обличчя з оригінальних зображень, і роздільна здатність обрізаного зображення обличчя становить 28 × 28 пікселів. Категорії емоцій на зображеннях у SAMM поділяються на категорії, включаючи «огиду», «страх», «презирство», «злість», «репресії», «здивування», «щастя» та «інші». Після класифікації за трьома категоріями емоцій кількість «негативних», «позитивних» і «сюрпризів» становить 92, 26,15.

CASME II [35]: CASME II включає 24 предмети та 145 зразків, які відповідають 145 емоціям. Усі зразки зняті лабораторними камерами з частотою кадрів 200. Оригінальний розмір зразків 640 × 480 пікселів. Роздільна здатність обрізаного та зміненого зображення обличчя становить 28 × 28 пікселів. Зразки в CASME II розділені на такі категорії, як «щастя», «здивування», «огида», «сум», «страх», «репресії» та «інші». Після об'єднання в три категорії емоцій кількість «негативних», «позитивних» і «сюрпризів» становить 88, 32,25. Початок, зміщення та індекс вершини анотовані в CASME II.

CASME III [36]: База даних спонтанного мікровиразу обличчя, відома як CASME III, знаходиться в третьому поколінні. З метою розпізнавання мікровиразів у CASME III, він пропонує глибоку інформацію та високу екологічну валідність. CASME III part A включає 100 предметів і 943 зразки, що відповідають 943 емоціям. Лабораторна камера знімає всі зразки з частотою кадрів 30 з вихідною роздільною здатністю 1280 × 720 пікселів. Індекси початку, зміщення та вершини наведено в CASME III, частина A. Ми використовуємо MTCNN для виявлення облич і кадрування зображень облич у 28 × 28 пікселів. Зразки в частині A CASME III класифікуються на «щастя», «гнів», «страх», «відраза», «здивування», «інші» та «сум». Загальна кількість «негативу», «позитиву» та «сюрпризу» становить 508, 64 та 201.

SMIC[34]: SMIC-HS складається з 16 предметів і 164 зразків, що відповідають 164 окремим емоціям. Усі зразки знімає лабораторна камера з частотою кадрів 100. Початковий розмір зображення зразків становить 640 × 480 пікселів. Розмір зображення обрізаного обличчя становить 28 × 28 пікселів. Зразки в SMIC класифікуються на «негативні», «несподівані» та «позитивні». Числа «негативний», «позитивний» і «сюрприз» — 70, 51 і 43. Початок і зсув наведено в SMIC, але індекс вершини не вказано в SMIC. Детальну інформацію про ці три набори даних можна узагальнити в таблиці 1.

Таблиця 1

Експерименти реалізовано в наборах даних SAMM[33], SMIC[34], CASME II[35] і CASME III [36]. SAMM, SMIC і CASME II об'єднані в один складений набір даних, і однакові мітки в цих трьох наборах даних застосовуються для завдань мікровиразу

Набори данних	SAMM	CASME II	SMIC	CASME III
Об'єкти	28	24	16	100

Приклади	133	145	164	943
Частота кадрів	200	200	100	30
Роздільна здатність обрізаного кадру	28 × 28	28 × 28	28 × 28	28 × 28
Негатив	92	88	70	508
Позитив	26	32	51	64
Сюрприз	15	25	43	201
Індекс початку	+	+	+	+
Індекс зміщення	+	+	+	+
Індекс вершини	+	+	-	+

Деталі реалізації

По-перше, оскільки в наборі даних SMIC відсутній піковий кадр, ми застосовуємо техніку D&C-RoIs [31] для визначення індексу пікового кадру. Для наборів даних SAMM, CASME II і CASME III надається основна правда для пікового кадру. Після отримання зображень початку та вершини з набору даних алгоритм Гуннара Фарнебека [37] використовується для виділення оптичного потоку із зображень на початку та піку. Розмір трьох елементів зображень оптичного потоку — горизонтального, вертикального та оптичного деформаційного елемента — змінено до 28 × 28 × 3 пікселів. За допомогою зображень оптичного потоку можна приблизно оцінити ступінь деформації обличчя. Після цього ми використовуємо MTCNN [32] для отримання координат орієнтирів обличчя із зображень вершини. Використовуючи орієнтири обличчя, ми виділяємо чотири важливі області обличчя: ліве око, ліва губа, праве око та права губа. Ці чотири карти характеристик оптичного потоку мають половину розміру всього зображення оптичного потоку, маючи розмір 14 × 14 × 3. Після цього наша мережа HTNet отримує комбіновані чотири функції оптичного потоку. Усі експерименти проводяться з використанням Pytorch і Python 3.9 на Ubuntu 22.04. Ми використовуємо швидкість навчання 5 × 10−5 для параметрів навчання з верхньою межею 800 епох.

Налаштування: стандартним методом оцінювання завдань мікровиразу є залишення одного суб'єкта (LOSO). Усі наші експерименти проводяться за допомогою перехресної перевірки LOSO для справедливого порівняння. Зразки одного суб'єкта зберігаються, а зразки інших суб'єктів використовуються для навчання. Ця техніка перехресної перевірки LOSO може моделювати реальну ситуацію, коли люди зареєстровані в різних середовищах і місцях, але походять з різних культур. Показники продуктивності. Розподіл класів у складеному наборі даних мікровиразів є незбалансованим. Коефіцієнт несподіванки: позитивний: негативний становить приблизно 1:1,3:3. Ми також використовуємо UAR та UF1 для звітування про наші результати, щоб врахувати ці дисбаланси класів.

1) Незважений F1-оцінка (UF1): Макро-середня F1-оцінка є іншою назвою для цього показника. Незважений показник F1, який може надати рівний пріоритет рідкісним класам, є корисним варіантом для оцінки незбалансованого мультикласу. Ми повинні з'ясувати всі помилкові позитивні результати (FP), справжні позитивні результати (TP) і помилкові негативні результати (FN) для всіх складок (LOSO) у кожному класі c, щоб отримати UF1. Для визначення UF1 можна усереднити F1-бали F1c для кожного класу:

$$F1_c = \frac{2 \times TP_c}{2 \times TP_c + FP_c + FN_c}$$
$$UF1 = \frac{F1_c}{C} \quad (8)$$

У формулі 8, C — кількість класів.

2) Незважене середнє запам'ятовування (UAR): UAR, яке тісно пов'язане зі стандартною точністю, є ще одним корисним показником для оцінки ефективності моделі, коли співвідношення класів вибірки є незбалансованим. При використанні середньозваженого запам'ятовування або стандартної точності в класах з дисбалансом, точна оцінка більших розмірів класу може бути необ'єктивною. TPc — це точність у кожному класі емоцій, C — кількість класів емоцій, а nc — загальний розмір вибірки для c-го класу.

$$UAR = \frac{1}{C}\sum_c \frac{TP_c}{n_c} \quad (9)$$

Обидва показники можна використовувати для оцінки ефективності моделі, яка однаково добре оцінює всі категорії. Ці показники можуть уникнути того, щоб деякі моделі добре передбачали лише певні класи.

Порівняння з сучасними досягненнями

У таблицях 2 і 3 ми повідомили про наш метод із попередніми ручними методами та методами глибокого навчання на наборах даних мікровиразу — CASME II, CASME III, SMIC і SAMM. Повідомляються результати незваженого показника F1 (UF1) і незваженого середнього

запам'ятовування (UAR). Жирний текст позначає найкращий результат. LBP-TOP [19]: вони запропонували використовувати локальні бінарні шаблони з шістьма точками перетину (LBP-SIP) для виділення рис обличчя. Запропонований метод LBP-SIP може ефективно зменшити надмірність у шаблонах LBP-TOP і зменшити обчислювальну складність. Крім того, вони використовують піраміду Гауса з кількома роздільними здатностями для виділення різних функцій роздільної здатності зображення та об'єднання всіх функцій для розпізнавання мікровиразів. Нарешті, вони використали класифікатор SVM із перехресною перевіркою LOSO на базових наборах даних мікровиразів і досягли гарної продуктивності.

Bi-WOOF [5]: звичайні підходи до виділення ознак передбачають використання всіх відеопослідовностей для завдань мікровиразу. Однак уся послідовність може містити шумову інформацію, а деякі кадри не важливі. Вони запропонували використовувати лише два кадри – зображення початку та зображення вершини для завдань мікровиразу. Сила виражених змін максимальна в апексній рамці. Тож вони запропонували використати двозважений оптичний потік, щоб захопити основні риси обличчя зображення верхівки. Запропонований ними підхід досяг кращої продуктивності. OFF-ApexNet [2]: оскільки впровадження оптимальних методів виділення ознак для тонких рухів виразів обличчя є складним, і багато підходів виділяють особливості непомітних рухів виразів обличчя за допомогою ручних рис. Для завдань мікровиразу вони запропонували використовувати поля оптичного потоку від початкового та пікового кадрів. У полях оптичного потоку горизонтальні та вертикальні характеристики отримуються та передаються в мережу на основі CNN для подальшого вдосконалення функцій. Після цього витягнуті функції з OFF-ApexNet будуть використані для класифікації мікровиразів.

STSTNet [12]: вони запропонували використати три поверхневу модель на основі CNN для отримання дискримінаційного представлення високого рівня для класифікації мікровираження емоцій. У мережі на основі CNN він включає згортковий рівень, повністю зв'язаний рівень і рівень об'єднання. Мережа на базі CNN виділить характеристики руху м'язів у двох напрямках: горизонтальні та вертикальні поля оптичного потоку та оптичне напруження. Після цього шар softmax буде використовуватися для визначення станів мікровиразу.

Dual-Inception [39]: щоб подолати проблеми розпізнавання мікровиразів між базами даних, вони запропонували використовувати дві початкові мережі для вилучення горизонтальних і вертикальних ознак у оптичній карті ознак. Їхній метод досяг кращої продуктивності в комбінованому базовому наборі даних мікроекспресії. Горизонтальний компонент оптичного потоку буде подаватися в одну початкову мережу, а вертикальний компонент оптичного потоку буде подаватися в іншу початкову мережу. Після цих двох початкових мереж об'єднана мережа об'єднає ці дві незалежні функції для надійної класифікації. EMR [4]: враховуючи, що набір даних про емоції обличчя має кінцеву кількість навчальних зразків, необхідно підвищити як кількість, так і якість навчальних зображень у наборі даних, EMR пропонує використовувати дві стратегії адаптації домену для покращення кількості навчальних зображень. Змагальні методи навчання можуть додавати більше згенерованих зразків у навчальні набори, а метод збільшення виразу може збільшити мікровираз одного зображення та полегшити мережеве вилучення корисних функцій. RCN [13]: у композитному наборі даних необхідний тонкий рух обличчя може зникнути через зміщення домену та погіршити продуктивність моделі. Таким чином, вони запропонували використовувати зображення меншого розміру як вхідні дані з моделлю меншої архітектури, яка може допомогти підвищити продуктивність моделі в задачах складених наборів даних. У своїй статті вони запропонували рекурентну згортку (RCN) для вилучення рис обличчя за допомогою трьох безпараметричних модулів — блоку уваги, швидкого підключення та широкого розширення. Нарешті, стратегія автоматичного пошуку використовується для оптимізації трьох параметрів.

FeatRef [14]: у мережі FeatRef вона включає два етапи: на першому етапі горизонтальна початкова мережа та вертикальна початкова мережа вилучатимуть горизонтальні та вертикальні характеристики руху м'язів. Після цього горизонтальні та вертикальні об'єкти будуть об'єднані та подані до трьох мереж на основі уваги для класифікації цих виділених об'єктів у різні категорії мікровиразу. Нарешті, класифікаційна гілка використовується для злиття помітних і розрізнювальних ознак, отриманих із початкового модуля для висновку про мікровираз. Їхні експерименти продемонстрували, що функція Feature Refinement (FeatRef) може виділити дискримінаційне та помітне представлення для розпізнавання мікровиразів і досягла хорошої продуктивності в наборах даних мікровиразів.

Порівняно з методами ручної роботи

У таблиці 2 LBP-TOP використовує локальну двійкову інформацію з шістьма точками перетину для виділення рис обличчя, що є ознакою на основі зовнішності. Навпаки, Bi-WOOF використовує функції оптичного потоку як вхідні дані, які є геометричними характеристиками. Обидва вони використовують SVM як класифікатор. Порівняно з Bi-WOOF, наш метод може покращити UF1 і UAR у композитних наборах даних з 0,6296 до 0,8603 і 0,6227 до 0,8475, що є покращенням більш ніж на 20%. Продуктивність моделі можна безперервно підвищувати за допомогою HTNet на CASME II, SMIC і SAMM із великим відривом, ніж ручні методи – LBP-TOP і Bi-WOOF. У таблиці 2 більшість методів

глибокого навчання можуть перевершити методи ручної роботи. Однак набір даних є складеним і має зміщення домену від одного набору даних до іншого. Деякі більш глибокі методи працюють гірше, ніж методи ручної роботи, такі як GoogLeNet і VGG.

Більш глибока мережа може переналаштовувати доменний шум і важко вивчати корисні функції вираження. Порівнюючи ALexNet, GoogLeNet і VGG16, він перевіряє, що дрібніші мережі можуть працювати краще, ніж глибші мережі, оскільки ALexNet є дрібнішою мережею, ніж GoogLeNet і VGG16.

Таблиця 2

Показники показника F1 (UF1) і UAR методів ручної роботи, методів глибокого навчання та нашого методу HTNet за протоколом LOSO. Жирний текст позначає найкращий результат.

Підхід	Повний		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP[19]	0,5882	0,5785	0,2	0,528	0,7026	0,7429	0,3954	0,4102
Bi-WOOF[5]	0,6296	0,6227	0,5727	0,5829	0,7805	0,8026	0,5211	0,5139
AlexNet[11]	0,6933	0,7154	0,6201	0,6373	0,7994	0,8312	0,6104	0,6642
GooLeNet[10]	0,5573	0,6049	0,5123	0,5511	0,5989	0,6414	0,5124	0,5992
VGG16[9]	0,6425	0,6516	0,58	0,5964	0,8166	0,8202	0,487	0,7493
OFF-ApexNet[2]	0,7196	0,7096	0,6817	0,6695	0,8764	0,8681	0,5409	0,5392
STSTNet[12]	0,7353	0,4905	0,6801	0,7013	0,8382	0,8686	0,6588	0,681
CapsuleNet[38]	0,652	0,6506	0,582	0,5877	0,7068	0,7018	0,6209	0,5959
Dual-Inception[39]	0,7322	0,7278	0,6645	0,6726	0,8621	0,856	0,5868	0,5663
EMR[4]	0,7885	0,7824	0,7461	0,753	0,8293	0,8209	0,7754	0,7152
RCN[13]	0,7432	0,719	0,6326	0,6441	0,8512	0,8123	0,7601	0,6715
FeatRef[14]	0,7838	0,7832	0,7011	0,7083	0,8915	0,8873	0,7372	0,7155
HTNet(Обраний)	**0,8603**	**0,8475**	**0,8049**	**0,7905**	**0,9532**	**0,9516**	**0,8131**	**0,8124**

Таблиця 3

Ефективність показника F1 (UF1) і UAR методів глибокого навчання та нашого методу HTNet за протоколом LOSO на CASME III Частина A. Жирним шрифтом позначено найкращий результат.

Підходи	CASME III Частина A	
	UF1	**UAR**
AlexNet [11]	0.257	0.2634
STSTNet [12]	0.3795	0.3792
RCN [13]	0.3928	0.3893
FeatRef [14]	0.3493	0.3413
HTNet(Обраний метод)	**0.5767**	**0.5415**

Порівняно з методами глибокого навчання

У таблицях 2 і 3 наш HTNet значно перевершує більшість методів глибокого навчання. Згідно з таблицею 2, HTNet досяг UF1 і UAR 0,8603 і 0,8475 у повних складених трьох наборах даних, що становить приблизно 7% покращення порівняно з попередніми методами sota. У таблиці 2 ми бачимо, що ALexNet, GoogLeNet і VGG16 досягли UF1 і UAR 0,6933 і 0,7154, 0,5573 і 0,6049, 0,6425 і 0,6516 у повному сукупному наборі даних. Вони досягли гірших результатів, ніж інші методи глибокого навчання в повних складених наборах даних. Причина полягає в тому, що ці три глибокі методи використовують більш глибоку мережу, яка легше вивчає інформацію про шум через обмежену кількість навчальних зразків. OFF-ApexNet, STSTNet, CapsuleNet і Dual-Inception використовують оптичний потік як вхідні дані та більш дрібну мережу для вилучення рухів обличчя. Вони досягли кращих результатів, ніж ALexNet, GoogLeNet і VGG16 у повних складених наборах даних. Однак їхні мережі не включають методи збагачення даних і не мають блоків уваги, тому їхні мережі можуть досягати неоптимальних результатів. Порівняно з OFF-ApexNet, STSTNet, CapsuleNet і Dual-Inception, EMR використовує методи збагачення даних і посилює мікровираз, завдяки чому мережі легше вивчати функції руху обличчя. Отже, EMR працює краще, ніж OFF-ApexNet, STSTNet, CapsuleNet і Dual-Inception, що має приблизно 5% покращення в UF1 і UAR повних складених наборів даних. Однак EMR не включає блоки уваги до своїх мереж; ця мережа не може зосередитися на важливих областях обличчя, тому EMR не може отримати кращий результат, ніж RCN в UF1 CASME II. Хоча додати більше навчальних зразків у навчальні набори. Порівняно з EMR і RCN, FeatRef використовує дві початкові мережі для виділення характеристик горизонтальних компонентів і характеристик

вертикальних компонентів оптичного потоку. Після цього об'єкти горизонтального компонента та об'єкти вертикального компонента будуть об'єднані разом і подані в мережу навчання властивостей виразу, яка включає блок уваги. Блок уваги може автоматично регулювати ваги та зосереджуватися на цих двох важливих функціях. FeatRef досяг більш конкурентоспроможних результатів, ніж EMR і RCN. Однак FeatRef реалізує самоконтроль на всіх картах функцій оптичного потоку; під час обчислення карт характеристик оптичного потоку це вносить інформацію про шум. Деяка деформація пікселів може насправді не відображати рухи м'язів на краю карт функцій оптичного потоку. Ця деформація пікселів, зазначена на картах функцій оптичного потоку, може бути внесена камерою. Крім того, реалізація самоконтролю на всій карті функцій оптичного потоку не може охопити дрібнозернисті та крупнозернисті характеристики в різних роздільних здатностях карт функцій оптичного потоку. HTNet зосереджується на чотирьох зонах обличчя замість усієї області обличчя: області лівого ока, області правого ока, області лівих губ і області правих губ, що може усунути звуковий фоновий шум, який уловлює лабораторна камера через потенційне мерехтіння світла. З локальною самоувагою в кожній області обличчя трансформерний шар можна використовувати для зосередження на локалізації помірного скорочення м'язів. Крім того, рівень агрегації для вивчення взаємодії між різними роздільними здатностями карт функцій оптичного потоку. Отже, HTNet досяг кращої продуктивності, ніж попередні методи на повних складених наборах даних.

У таблиці 3 ми проводимо експерименти з CASME III, частина A, і повідомляємо про незважену оцінку F1 (UF1) і незважене середнє запам'ятовування (UAR) для методів глибокого навчання та нашого методу HTNet за протоколом LOSO. HTNet перевершує попередні методи за конкурентною перевагою. Крім того, результат показує, що RCN працює краще, ніж FeatRef, при цьому показник UF1 зріс з 0,3493 до 0,3928, а UAR — з 0,3413 до 0,3893. Хоча в таблиці 2 FeatRef працює краще, ніж RCN. Причина полягає в тому, що RCN використовує методи пошуку параметрів для оптимізації вільних параметрів для отримання кращих представлень, одночасно запобігаючи переобладнанню шляхом зменшення складності навчання.

Апробаційне дослідження

У цьому розділі розглядається вплив розміру блоку, прихованого розміру, кількості головок у трансформері та різної кількості шарів трансформера на кожному рівні ієрархії. Базові експериментальні налаштування використовують розмір блоку 7 × 7 на нижньому рівні, прихований розмір — 256, кількість головок у трансформері — 3, а кількість шарів трансформера на кожному рівні ієрархії — (2, 2, 8). Кожен експеримент з абляції змінюватиме кожен параметр окремо, хоча в інших експериментах використовуватимуться ті самі основні налаштування. Метрика UAR і метрика UF1 використовуються для оцінки кожного з цих підходів.

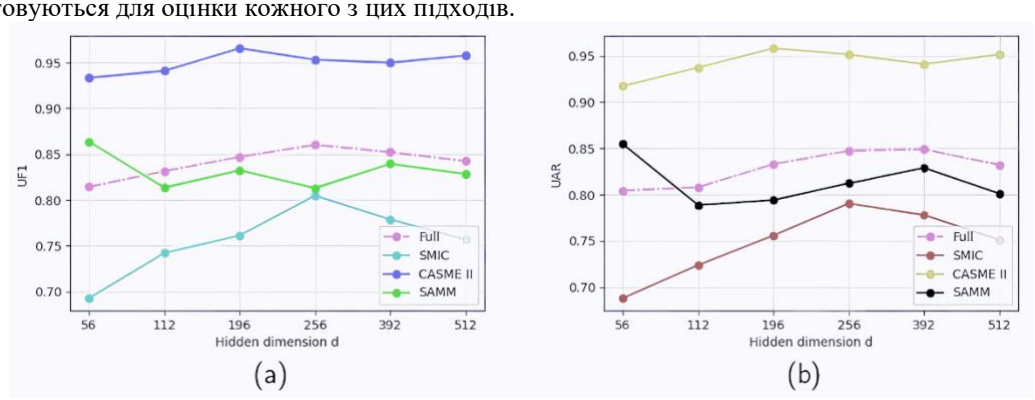

Рис. 5. Ми досліджуємо, як кількість вимірів впливає на точність складених наборів даних – SMIC, SAMM і CASME II. Повідомляється про ефективність незваженого показника F1 (UF1) і незваженого середнього запам'ятовування (UAR) складених наборів даних

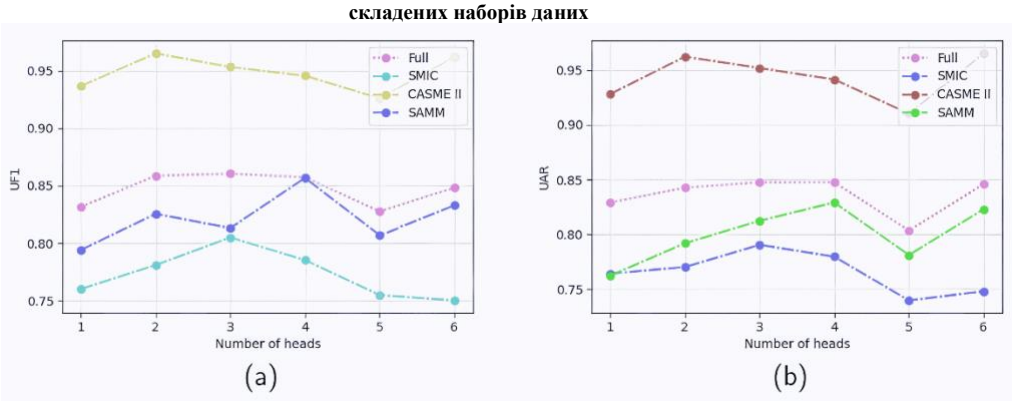

Рис. 6. Ми досліджуємо вплив підрахунку напору трансформерного рівня на точність у складених наборах даних SMIC, SAMM і CASME II. Повідомляється про ефективність незваженого показника F1 (UF1) і незваженого середнього запам'ятовування (UAR) складених наборів даних

Вплив розміру блоку

Ми розуміємо, що різні розміри блоків впливатимуть на загальну точність складених наборів даних мікровиразів. Більший розмір блоку вказує на те, що чотири області обличчя більше перекриватимуться. Тому ми проводимо різні експерименти, щоб вивчити вплив кількості розмірів блоків на складені набори даних – SMIC, SAMM і CASME II. Результати представлені в таблиці 4; розмір блоку нижнього рівня збільшується з 5 до 10, а розмір блоку середнього рівня в 2 рази перевищує розмір блоку нижнього рівня. Розмір блоку верхнього рівня в 4 рази перевищує розмір блоку нижнього рівня. Менший розмір блоку матиме гіршу продуктивність, оскільки відсутні деякі з необхідних лицьових частин. Більший розмір блоку працюватиме краще, ніж менший розмір блоку; однак, якщо розмір блоку стане занадто великим, це вплине на продуктивність моделі. Причина полягає в тому, що більший розмір блоку вказує на те, що деякі з цих чотирьох областей обличчя будуть перекриватися, наприклад, область лівого ока накладається на область правого ока. Перекриття зашкодить продуктивності моделі.

Таблиця 4

Вивчення впливу різних розмірів блоків на складені набори даних-SMIC, SAMM і CASME II. Повідомляється про ефективність показника F1 (UF1) і UAR складених наборів даних.

Розмір блоку (Нижній рівень)	5x5	6x6	7x7	8x8	9x9	10x10
Розмір блоку (Середній рівень)	10x10	12x12	14x14	16x16	18x18	20x20
Розмір блоку (Верхній рівень)	20x20	24x24	28x28	32x32	36x36	40x40
Повна UF1	0,837	0,8271	**0,8603**	0,85	0,8511	0,8546
Повна UAR	0,8213	0,8037	0,8475	**0,846**	0,7445	0,8383
SMIC UF1	0,7556	0,7394	**0,8049**	0,7833	0,7753	0,8008
SMIC UAR	0,7469	0,7214	**0,7905**	0,7792	0,7708	0,7892
SAMM UF1	**0,8237**	0,7903	0,8131	0,7909	0,8168	0,8068
SAMM UAR	0,8033	0,775	**0,8124**	0,79	0,8088	0,7812
CASME II UF1	0,9422	0,9482	0,9532	**0,964**	0,9722	0,957
CASME II UAR	0,9308	0,9317	0,9516	0,962	**0,9658**	0,945

Впливи розмірів

Ми досліджуємо, як кількість вимірів впливає на точність складених наборів даних - SMIC, SAMM і CASME II. Результати незваженого показника F1 (UF1) і незваженого середнього запам'ятовування (UAR) складених наборів даних представлені на рис.5. Менший прихований вимір має гіршу продуктивність, оскільки малий прихований вимір важко закодувати на карті функцій оптичного потоку. Однак використання більшого прихованого розміру призведе до проблем із переобладнанням. Повна оцінка UF1 для прихованого виміру 256 має найкращу продуктивність, близько 0,86.

Вплив кількості голів

Ми досліджуємо вплив підрахунку напору трансформерного рівня на точність у складених наборах даних – SMIC, SAMM і CASME II. Результати незваженого показника F1 (UF1) і незваженого середнього запам'ятовування (UAR) складених наборів даних представлені на рис.6. На рис. 6 (а) показано, що три головки в трансформерних шарах матимуть найкращу продуктивність.

Вплив різної кількості трансформерних блоків

У таблиці 5 ми досліджуємо, як кількість трансформерних шарів впливає на точність, оцінену за допомогою кількох наборів даних. Повідомляється про показники незваженого показника F1 (UF1) і незваженого середнього запам'ятовування (UAR) складених наборів даних. У нашій моделі перші два шари використовуються для виділення дрібнозернистих характеристик. Оскільки розміри блоків двох верхніх рівнів досить скромні, ми використовуємо меншу кількість трансформерних шарів у перших двох шарах, щоб досягти кращої продуктивності. Наша HTNet має однакову кількість трансформерних шарів на нижньому та середньому рівнях, тобто 2. Ми збільшуємо кількість трансформерних шарів верхнього рівня з 2 до 12. Це демонструє, що параметри та час навчання зростають разом із кількістю трансформерних шарів. Менша кількість трансформерних шарів матиме обмежену здатність кодувати характеристики оптичного потоку. Для порівняння, складність нашої моделі HTNet зросте через більшу кількість трансформерних шарів. Крім того, продуктивність моделі буде погіршуватися через проблеми з переобладнанням, спричинені збільшенням кількості шарів трансформера.

Таблиця 5

Дослідження впливу кількості шарів трансформера на складені набори даних. Повідомляється про ефективність показника F1 (UF1) і UAR складених наборів даних.

Кількість рівнів трасформера	(2,2,2)	(2,2,4)	(2,2,6)	(2,2,8)	(2,2,10)	(2,2,12)
Повна UF1	0,8105	0,7297	0,8316	**0,8603**	0,8464	0,8499
Повна UAR	0,7848	0,8029	0,8183	**0,8475**	0,8376	0,8207
SMIC UF1	0,748	0,7249	0,744	**0,8049**	0,76	0,7561
SMIC UAR	0,7363	0,7065	0,734	**0,7905**	0,756	0,7345
SAMM UF1	0,7984	7137	0,7732	0,8131	0,8571	**0,8682**
SAMM UAR	0,75	0,7849	0,7742	0,8124	**0,8453**	0,8419
CASME II UF1	0,8845	0,9647	**0,9722**	0,9532	0,9492	0,95
CASME II UAR	0,8588	0,9554	**0,9658**	0,9516	0,9346	0,9412
Параметри(мільйони)	149,57	245,88	342,2	438,51	534,82	631,13
Час тренування(с)	3767	4836	5906	6942	8085	9147

Якісний та кількісний аналіз запропонованого методу

На рис.7 (a) показані матриці плутанини HTNet. HTNet досягає точності 0,94, 0,81, 0,80 для негативних, позитивних і несподіваних, відповідно, у повних складених наборах даних. У цих SMIC, SAMM і CASME II негативна категорія отримує найвищу точність, оскільки в цих трьох наборах даних негативні зразки мають найбільше навчальне число, і нашу модель можна добре навчити. У той час як у SMIC і SAMM наша модель іноді не може розрізнити категорії подиву та негативних емоцій, оскільки немає достатньо навчальних зразків. CASME II забезпечує більш точний апекс-кадр, тож ми можемо отримати точніші карти функцій оптичного потоку, які краще визначають рухи м'язів обличчя. У CASME II негатив, позитив і сюрприз мають відносно високу точність, тобто понад 90%. Лише невелика частина тестових зразків неправильно класифікована. Крім того, SMIC використовує нижчу частоту кадрів для захоплення зображення. Фоновий шум має більший вплив на них, як-от мерехтіння світла, тіні та освітлення під час обчислення карт функцій оптичного потоку, тому позитивні та несподівані дані в наборі даних SMIC досягли 76% і 72% точності.

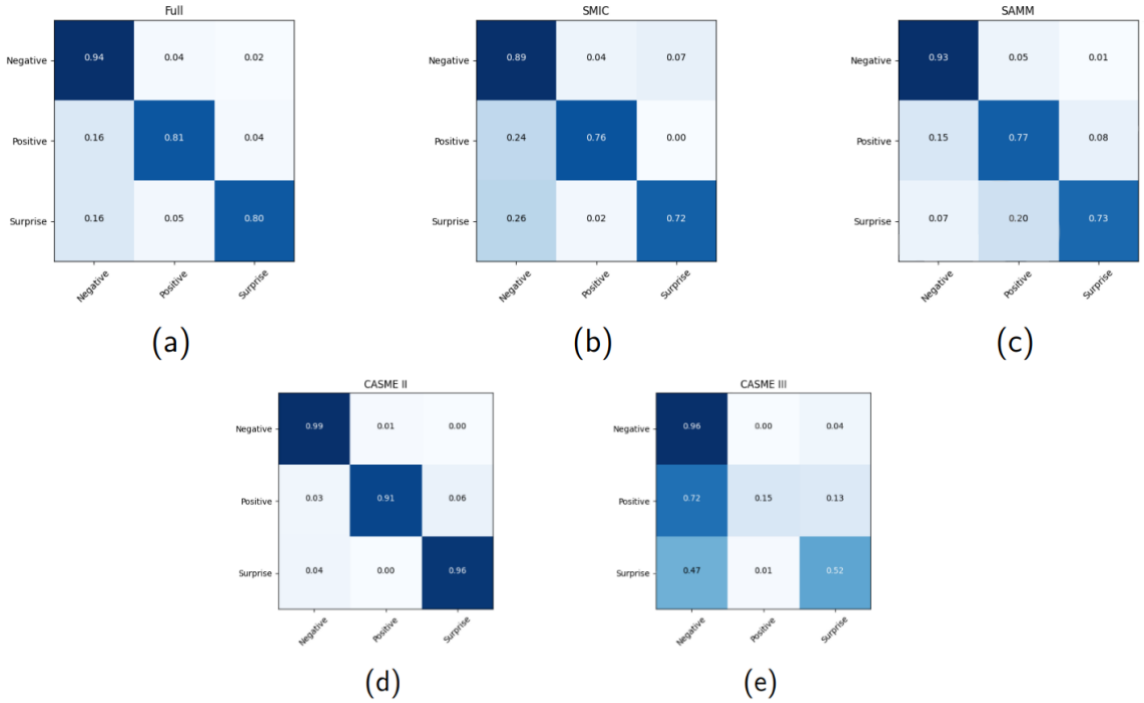

Рис. 7: Матриця плутанини для запропонованої HTNet у зведеній базі даних – SMIC, SAMM і набір даних CASME II і CASME III з використанням 3 класів

Ми впроваджуємо перехресну перевірку без одного суб'єкта в CASME III. CASME III містить більше зразків, ніж інші набори даних. На рис. 7 (e) наведено матрицю помилок для CASME III. Матриця плутанини демонструє, що негатив досяг найкращої точності порівняно з двома іншими категоріями емоцій. Позитивний досяг найгіршої точності, яка становить близько 0,15. Більшість позитивних зразків неправильно класифікуються як негативні. Хоча в складених наборах даних лише невелика частина зразків помилково класифікується як негативні. Оскільки CASME III відрізняється

від попередніх наборів даних, він містить більш складну інформацію щодо розподілу типів вибірки та осіб.

Висновок

У цьому документі ми пропонуємо ієрархічну архітектуру трансформера для виділення чотирьох важливих ознак обличчя для розпізнавання мікровиразів. Порівняно з попередніми методами глибокого навчання, однією з ключових сильних сторін моделі HTNet є її здатність моделювати довгострокові тимчасові залежності, що має вирішальне значення для точного розпізнавання мікровиразів, які часто виникають протягом короткого періоду часу. Ієрархічна архітектура моделі дозволяє виділяти ознаки в кількох часових масштабах, що може підвищити точність розпізнавання, фіксуючи як короткострокову, так і довгострокову динаміку виразу. Крім того, модель HTNet базується на архітектурі Transformer, яка продемонструвала найсучаснішу продуктивність у задачах обробки природної мови. Ця архітектура добре підходить для розпізнавання мікровиразів, оскільки дозволяє моделювати складні залежності між різними частинами виразу обличчя, а також є ефективною з точки зору обчислень. Незважаючи на ці сильні сторони, модель HTNet вимагає великої кількості навчальних даних, які може бути важко отримати в деяких сценаріях. Модель також може бути чутливою до змін освітлення, пози та інших факторів навколишнього середовища, які можуть вплинути на появу мікровиразів. Щоб усунути ці обмеження, майбутня робота може вивчити шляхи підвищення ефективності та надійності моделі HTNet. Методи розширення даних можна використовувати для збільшення обсягу навчальних даних і покращення здатності моделі узагальнювати різні умови середовища. Трансферне навчання також можна використовувати для використання попередньо навчених моделей для пов'язаних завдань і підвищення ефективності моделі HTNet. З подальшим розвитком і оптимізацією модель HTNet може виявитися цінним інструментом для точного розпізнавання мікровиразів у різних програмах, включаючи детекцію брехні, розпізнавання емоцій і оцінку психічного здоров'я.

Література

1. Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, W. Zhang, Ferv39k: A large-scale multiscene dataset for facial expression recognition in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20922–20931.
2. Y. S. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, L.-K. Tan, Off-apexnet on micro-expression recognition system, Signal Processing: Image Communication 74 (2019) 129–139.
3. Y. Liu, W. Wang, C. Feng, H. Zhang, Z. Chen, Y. Zhan, Expression snippet transformer for robust video-based facial expression recognition, Pattern Recognition (2023) 109368.
4. Y. Liu, H. Du, L. Zheng, T. Gedeon, A neural micro-expression recognizer, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–4.
5. S.-T. Liong, J. See, K. Wong, R. C.-W. Phan, Less is more: Micro-expression recognition from video using apex frame, Signal Processing: Image Communication 62 (2018) 82–92.
6. Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, X. Fu, A main directional mean optical flow feature for spontaneous micro-expression recognition, IEEE Transactions on Affective Computing 7 (4) (2015) 299–310.
7. S. Happy, A. Routray, Fuzzy histogram of optical flow orientations for micro-expression recognition, IEEE Transactions on Affective Computing 10 (3) (2017) 394–406.
8. S.-T. Liong, R. C.-W. Phan, J. See, Y.-H. Oh, K. Wong, Optical strain based recognition of subtle emotions, in: 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), IEEE, 2014, pp. 180–184.
9. A. Sengupta, Y. Ye, R. Wang, C. Liu, K. Roy, Going deeper in spiking neural networks: Vgg and residual architectures, Frontiers in neuroscience 13 (2019) 95.
10. P. Ballester, R. M. Araujo, On the performance of googlenet and alexnet applied to sketches, in: Thirtieth AAAI conference on artificial intelligence, 2016.
11. H. Zhang, H. Zhang, A review of micro-expression recognition based on deep learning, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 01–08.
12. S.-T. Liong, Y. S. Gan, J. See, H.-Q. Khor, Y.-C. Huang, Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–5.
13. Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, IEEE Transactions on Image Processing 29 (2020) 8590–8605.
14. L. Zhou, Q. Mao, X. Huang, F. Zhang, Z. Zhang, Feature refinement: An expression-specific feature learning and fusion method for micro-expression recognition, Pattern Recognition 122 (2022) 108275.
15. K. Gedamu, Y. Ji, L. Gao, Y. Yang, H. T. Shen, Relation-mining self-attention network for skeleton-based human action recognition, Pattern Recognition (2023) 109455.

16. W. Sheng, X. Li, Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network, Pattern Recognition 114 (2021) 107868.

17. G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence 29 (6) (2007) 915–928.

18. S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, X. Fu, Micro-expression recognition using dynamic textures on tensor independent color space, in: 2014 22nd international conference on pattern recognition, IEEE, 2014, pp. 4678–4683.

19. Y. Wang, J. See, R. C.-W. Phan, Y.-H. Oh, Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition, in: Asian conference on computer vision, Springer, 2014, pp. 525–537.

20. S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, C.-G. Zhou, Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features, in: European Conference on computer vision, Springer, 2014, pp. 325–338.

21. X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietik¨ainen, Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns, Neurocomputing 175 (2016) 564–578.

22. X. Li, J. Yu, S. Zhan, Spontaneous facial micro-expression detection based on deep learning, in: 2016 IEEE 13th International Conference on Signal Processing (ICSP), IEEE, 2016, pp. 1130–1134.

23. L. Lei, T. Chen, S. Li, J. Li, Micro-expression recognition based on facial graph representation learning and facial action unit fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1571–1580.

24. K. Zhang, Y. Huang, H. Wu, L. Wang, Facial smile detection based on deep learning features, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE, 2015, pp. 534–538.

25. Z. Xia, X. Hong, X. Gao, X. Feng, G. Zhao, Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions, IEEE Transactions on Multimedia 22 (3) (2019) 626–640.

26. H. Chen, R. Liu, Z. Xie, Q. Hu, J. Dai, J. Zhai, Majorities help minorities: Hierarchical structure guided transfer learning for few-shot fault recognition, Pattern Recognition 123 (2022) 108383.

27. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations.

28. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

29. J. Gan, Y. Chen, B. Hu, J. Leng, W. Wang, X. Gao, Characters as graphs: Interpretable handwritten chinese character recognition via pyramid graph transformer, Pattern Recognition (2023) 109317.

30. Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. O. Arik, T. Pfister, Nested hierarchical transformer: Towards ¨ accurate, data-efficient and interpretable visual understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 3417–3425.

31. S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, R. Phan, Automatic apex frame spotting in micro-expression database, in: 2015 3rd IAPR Asian conference on pattern recognition (ACPR), IEEE, 2015, pp. 665–669.

32. E. Jose, M. Greeshma, M. T. Haridas, M. Supriya, Face recognition based surveillance system using facenet and mtcnn on jetson tx2, in: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE, 2019, pp. 608–613.

33. A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE transactions on affective computing 9 (1) (2016) 116–129.

34. X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietik¨ainen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

35. W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, PloS one 9 (1) (2014) 86041.

36. J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang, X. Fu, Cas (me) 3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity, IEEE Transactions on Pattern Analysis & Machine Intelligence (01) (2022) 1–1.

37. C.-W. Chang, Z.-Q. Zhong, J.-J. Liou, A fpga implementation of farneback optical flow by high-level synthesis, FPGA '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 309.

38. N. Van Quang, J. Chun, T. Tokuyama, Capsulenet for micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–7.

39. L. Zhou, Q. Mao, L. Xue, Dual-inception network for cross-database micro-expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.