УДК 004.8

С. Л. Козлов; О. К. Колесницький, канд. тех. наук, проф.

ЗАСТОСУВАННЯ АРХІТЕКТУРИ ТРАНСФОРМЕР ДО ЗАДАЧІ SUPER-RESOLUTION

Протягом останніх 15-ти років згорткові нейронні мережі є основним підходом для вирішення задач комп'ютерного зору, і демонструють високий рівень продуктивності. Проте, архітектура трансформер, яка показала високі досягнення в галузі обробки природної мови, знаходить все ширше застосування до задач комп'ютерного зору і демонструє співставні або кращі результати. Нами розглянуто застосування архітектури трансформер до задачі super-resolution, а також наведено короткий огляд попередніх підходів. Безпосереднє застосування оригінальної архітектури трансформер дозволило забезпечити продуктивність, співставну з актуальними згортковими нейронними мережами. Проте, ефективне застосування архітектури трансформер до задач комп'ютерного зору пов'язане з викликами, які витікають з відмінностей між візуальним і мовленнєвим доменами. Перша відмінність - масштаб, оскільки зображення містять візуальні елементи різних масштабів, це ускладнює їх обробку за допомогою архітектури трансформер, що аналогічно до обробки токенів в ОПМ, працює з фрагментами одного розміру. Друга – об'єм інформації, адже обчислювальна складність обрахунку самоуваги квадратична довжині вхідної послідовності, що стає особливо критичним при обробці зображень високої роздільної здатності.

У статті проведено аналіз 12 робіт з цієї тематики, опублікованих починаючи з 2021 року, які пропонують підходи до усунення зазначених складнощів. В проаналізованих роботах можуть бути виділені наступні напрямки: дослідження застосування локальної уваги з вікнами різних форм, зокрема вікнами розрідженої уваги; дослідження канальної самоуваги та її поєднання з просторовою; дослідження можливості розширення архітектури трансформер за допомогою згорткових блоків. Означені дослідження дозволили суттєво збільшити якість відтворених зображень, проте не є вичерпними.

Ключові слова: *super-resolution, архітектура трансформер, згорткова нейронна мережа, комп'ютерний зір.*

Вступ

Швидкий розвиток технологій цифрової обробки зображень призвів до зростаючого попиту на зображення високої роздільної здатності у різноманітних сферах застосування, починаючи від медичної візуалізації та систем спостереження, закінчуючи виробництвом розважального мультимедійного контенту. Проте, отримання зображень високої роздільної здатності часто обмежується можливостями фоточутливих сенсорів, або іншими фізичними обмеженнями засобів захоплення, що призводить до зниження роздільної здатності і, як результат, деталізації і якості зображень. Super-resolution (SR) – процес відтворення зображення високої роздільної здатності з відповідного зображення низької роздільної здатності, який привернув останнім часом значну увагу як ефективний і дешевий спосіб вирішення цієї проблеми.

Традиційні методи SR, наприклад бікубічна інтерполяція, є простими і ефективними, але схильні до розмиття деталей і кільцевих артефактів, що негативно впливає на якість відновленого зображення. Вдосконалені методи, що ґрунтуються на навчанні, такі як методи розрідженого кодування або локальної лінійної регресії, були запропоновані для усунення цих недоліків. Швидке зростання обчислювальної потужності та широка доступність даних Big Date зробили можливим застосування глибокого навчання до задачі SR. Застосування згорткових та генеративно-змагальних нейронних мереж (ЗНМ та ГЗНМ) активно досліджувалось для вирішення задачі SR протягом останніх 10 років і дозволило досягти високого рівня якості відновлення і продемонструвало можливість адаптивності. Однак, незважаючи на досягнення ЗНМ, існують певні обмеження, пов'язані з властивістю локальності ЗНМ, що не дозволяє ефективно моделювати далекосяжні залежності, а також із

статичністю ваг згорткових фільтрів. ГЗНМ зосереджені на генерації зображень привабливих для ока, проте схильні до генерації артефактів та нестабільні в навчанні.

Архітектура трансформер нещодавно застосована до задач комп'ютерного зору високого рівня продемонструвала значний приріст продуктивності порівняно з ЗНМ. Трансформер, початково розроблений для задач обробки природної мови (ОПМ), спирається на механізм багатоголової самоуваги, який дозволяє безпосередньо моделювати далекосяжні залежності, аналізуючи взаємозв'язки між усіма елементами вхідного зображення. Проте обчислювальна складність такого підходу збільшується квадратично з розміром зображення, що ускладнює його застосування до задачі SR.

Метою статті є огляд і аналіз наявних підходів до вирішення задачі SR із застосуванням архітектури трансформер.

Задача super-resolution

Super-resolution – задача відновлення цифрового HR зображення (high resolution – висока роздільна здатність) з одного або декількох LR зображень (low resolution – низька роздільна здатність). По кількості вхідних LR-зображень поділяється на SISR (single-image super resolution) – одне вхідне LR зображення та MISR (multi-image super resolution) – декілька вхідних LR зображень. Переважна більшість досліджень фокусуються на SISR, через значно ширший спектр потенційних застосувань. Окрім того, техніки, досліджені для SISR, можуть бути застосовані для MISR. Нехай D – функція спотворення, що відображає зв'язок між LR-зображенням x і HR-зображенням y:

$$x = D(y, \delta), \#(1)$$

де δ – параметри функції спотворення, наприклад коефіцієнт зменшення або тип і рівень шуму. На практиці тип і параметри спотворення зазвичай невідомі, тому його моделюють, наприклад, за допомогою зменшення зображення бікубічною інтерполяцією. Тепер задачу SR можна визначити як пошук функції, оберненої до функції спотворення D, тобто необхідно знайти таку функцію M, що:

$$\hat{y} = M(x, \theta), \#(2)$$

де $\hat{y}$ – апроксимація початкового HR-зображення, θ – параметри функції M. Оскільки для одного LR-зображення може існувати декілька неідентичних відновлених HR-зображень, то задача SR є некоректно поставленою задачою.

Класифікацію наявних методів SISR наведено на рис. 1. Ранні методи SR базувались на застосування аналітичної інтерполяції, як то лінійна, бікубічна, інтерполяція кубічними сплайнами, чи New Edge Directed Interpolation [1]. Основною перевагою цих методів є простота і можливість застосування в реальному часі, проте просте правило інтерполяції призводить до значного розмиття деталей. Методи реконструкції [2, 3] використовують апріорні знання, для обмеження простору можливих рішень, що дозволяє генерувати чіткіші деталі. Однак, продуктивність багатьох методів реконструкції, швидко знижується зі збільшенням коефіцієнта масштабування, крім того ці методи є ресурсозатратними. Методи навчання у вирішенні задачі SISR отримали широку популярність через їх високу продуктивність та прийнятну обчислювальну складність. Тут застосовується машинне навчання для пошуку статистичних залежностей між фрагментами HR та LR зображень. Разом із розвитком машинного навчання, широке різноманіття моделей було застосовано до задачі SR: метод вкладення найближчих сусідів [4], методи розрідженого кодування [5], методи локальної лінійної регресії [6].

З початком розвитку глибокого навчання у 2012 [7] ЗНМ стали стандартом у вирішення задач комп'ютерного зору, зокрема і у задачах SR. Так, у SRCNN [8] запропоновано ЗНМ з трьох шарів, яка перевершила результати наявних методів навчання SR. Згодом її результат

було покращено завдяки збільшенню глибини мережі VDSR [9], або додаванню залишкових зв'язків SRResNet [10]. У ESDR [11] оптимізовано архітектуру SRResNet, ця мережа показала виняткові результати і стала еталонною для майбутніх досліджень. У ESPCN [12] запропоновано субпіксельну згортку завдяки якій стало можливим виконувати операцію підвищення дискретизації останнім кроком, що допомогло знизити вимоги по пам'яті та підвищити ефективність. У RCAN [13] запропоновано мережу ЗНМ з канальною увагою.

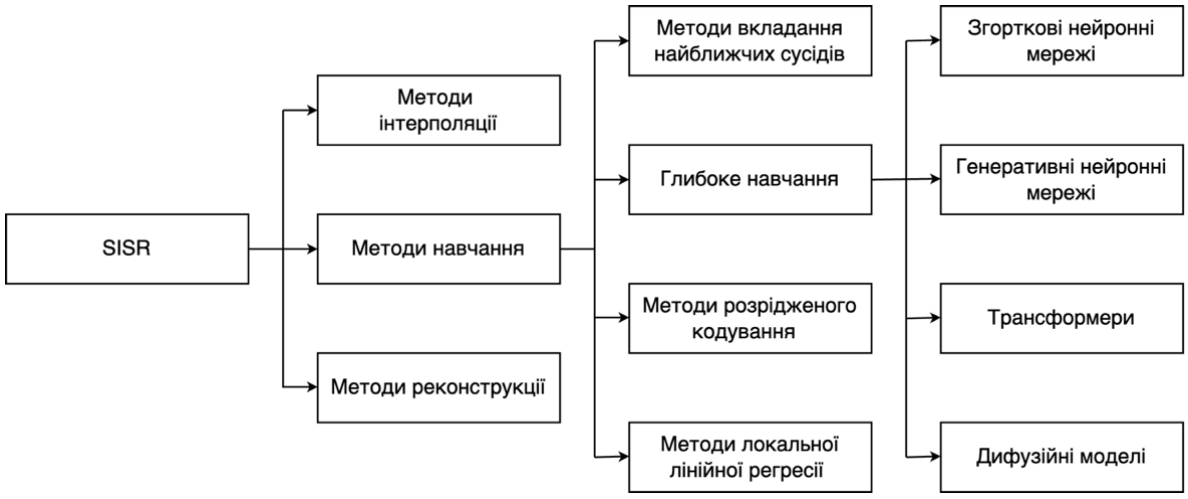

Рис. 1. Класифікація існуючих методів SR

Альтернативою до ЗНМ-методів є генеративні методи, а саме, методи, що базуються на ГЗНМ і дифузійних моделях. У SRGAN [10] запропоновано ГЗНМ мережу, яка завдяки поєднанню змагальної функції втрат і функції витрат вмісту дозволяє отримати відтворенні зображення вищої якості з точки зору сприйняття людиною. ESRGAN [14] є розвитком SRGAN і наразі є еталоном для ГЗНМ-методів. Дифузні моделі SRDiff [15] – відносно новий напрямок, що дозволяє надалі скоротити розрив між якістю відновленого зображення та суб'єктивним сприйняттям результату людиною, проте вимагає значних ресурсних затрат.

Починаючи з 2017 архітектура трансформер здійснила значний прорив у сфері ОПМ. Механізм самоуваги та нова архітектура мережі довели свою ефективність у обробці послідовних даних. Пізніше, у 2020-му, було запропоновано архітектуру ViT (Vision Transformer) [18], яка є адаптацію архітектури трансформер до задач комп'ютерного зору. ViT показала високу продуктивність і конкурентоспроможність в порівнянні з ЗНМ. Її застосування у сфері комп'ютерного зору активно досліджується, зокрема і у задачі SR.

У випадку методів навчання, задачу SR зводять до задачі оптимізації, тобто пошуку такого набору параметрів $\hat{\theta}$ функції M, який мінімізує значення функції втрат L для оригінального HR зображення y і його апроксимації $\hat{y}$:

$$\hat{\theta} = argmin_{\hat{\theta}} L(\hat{y}, y). \#(3)$$

Найпоширенішими функціями втрат є MAE (mean absolute error – середня абсолютна помилка) формула 5, та MSE (mean squared error – середньоквадратична помилка) формула 6. Однак, через високу чутливість MSE до аномальних значень, MAE частіше зустрічається у літературі. Також широко вживається функція втрат Charbonnier [16] формула 7. Для пари зображень y, $\hat{y}$ з шириною w, висотою h та кількістю каналів c визначимо кількість точок як $N_y = w \cdot h \cdot c$, а простір можливих позицій як:

$$\Omega_y = \{(i, j, k) \in \mathbb{N}_1^3 | i < h, \quad j < w, k < c\}, \#(4)$$

тоді, функції втрат можна означити як:

$$L_{MAE}(\hat{y}, y) = \frac{1}{N_y}\sum_{p\in\Omega_y}\left|y_p - \hat{y}_p\right|, \#(5)$$

$$L_{MSE}(\hat{y}, y) = \frac{1}{N_y}\sum_{p\in\Omega_y}\left|y_p - \hat{y}_p\right|^2, \#(6)$$

$$L_{\text{Charbonnier}}(\hat{y}, y) = \frac{1}{N_y}\sum_{p\in\Omega_y}\sqrt{\left|y_p - \hat{y}_p\right|^2 + \varepsilon^2}\ \#(7)$$

де $\varepsilon \in (0, 1]$ – константа, що гарантує відмінність підкореневого виразу від 0.

Оцінка якості відтвореного зображення є складною задачею, адже насамперед визначається людиною, яка його сприйматиме і залежить від багатьох властивостей, як то різкість, контраст або відсутність шуму. Очевидно, найкращий результат даватимуть методи, що спираються на суб'єктивну оцінку людиною, як наприклад MOS (mean opinion score – середня оцінка вражень). Проте, залучення людського ресурсу є часозатратним і обтяжливим, особливо для великих наборів даних. Альтернативою є застосування еталонних зображень, та проведення об'єктивної оцінки. Найбільш поширеною метрикою об'єктивної оцінки є PSNR (peak signal-to-noise ratio – пікове співвідношення сигнал/шум), яка є співвідношенням між максимальним рівнем сигналу L (256 для зображень з 8 біт/канал) та MSE для оригінального і відтвореного зображень:

$$PSNR(\hat{y}, y) = 10 \cdot log_{10}\frac{L^2}{\frac{1}{N_y}\sum_{p\in\Omega_y}|y_p - \hat{y}_p|}. \#(8)$$

Альтернативною метрикою, що краще відповідає вимогами оцінки зображень з точки зору сприйняття людиною, є SSIM (structural similarity index measure – індекс структурної подібності). SSIM ґрунтується на порівняльній оцінці трьох складових: яскравості, контрасту і структурної подібності [17]:

$$SSIM(\hat{y}, y) = \frac{\left(2\mu_{\hat{y}}\mu_y + c_1\right)\left(2\sigma_{\hat{y}y} + c_2\right)}{\left(\mu_{\hat{y}}^2 + \mu_y^2 + c_1\right)\left(\sigma_{\hat{y}}^2 + \sigma_y^2 + c_2\right)}, \#(9)$$

де, μ_y і $\mu_{\hat{y}}$ – середні значення яскравості пікселів оригінального і відтвореного зображень, σ_y і $\sigma_{\hat{y}}$ – середньоквадратичне відхилення яскравості пікселів оригінального і відтвореного зображень, $\sigma_{\hat{y}y}$ – коваріація, $c_i = (k_i L)^2$ – змінні, що запобігають діленню на 0, $k_1 = 0.01$, $k_2 = 0.03$.

Архітектура Visual Transformer

Запропонована у [18] архітектура ViT (Visual Transformer) є безпосередньою адаптацією архітектури трансформер, запропонованої у [19], до задач комп'ютерного зору. На рис. 2 зображені основні елементи ViT архітектури. Visual Transformer складається з N блоків, аналогічних блокам кодування оригінального трансформера. Кожен блок складається з двох послідовних підблоків з залишковими зв'язками: блоку багатоголової самоуваги та блоку повнозв'язної мережі прямого поширення.

Вхідне зображення представляють у вигляді вкладень фрагментів, аналогічно до вкладень токенів у випадку з ОПМ. Для цього його розбивають на частини, кожну з яких представляють у вигляді одновимірного вектора, вкладення фрагментів отримують за допомогою лінійного перетворення зазначених раніше одновимірних векторів. Альтернативним підходом до формування вкладень є застосування одного або декількох згорткових шарів [20].

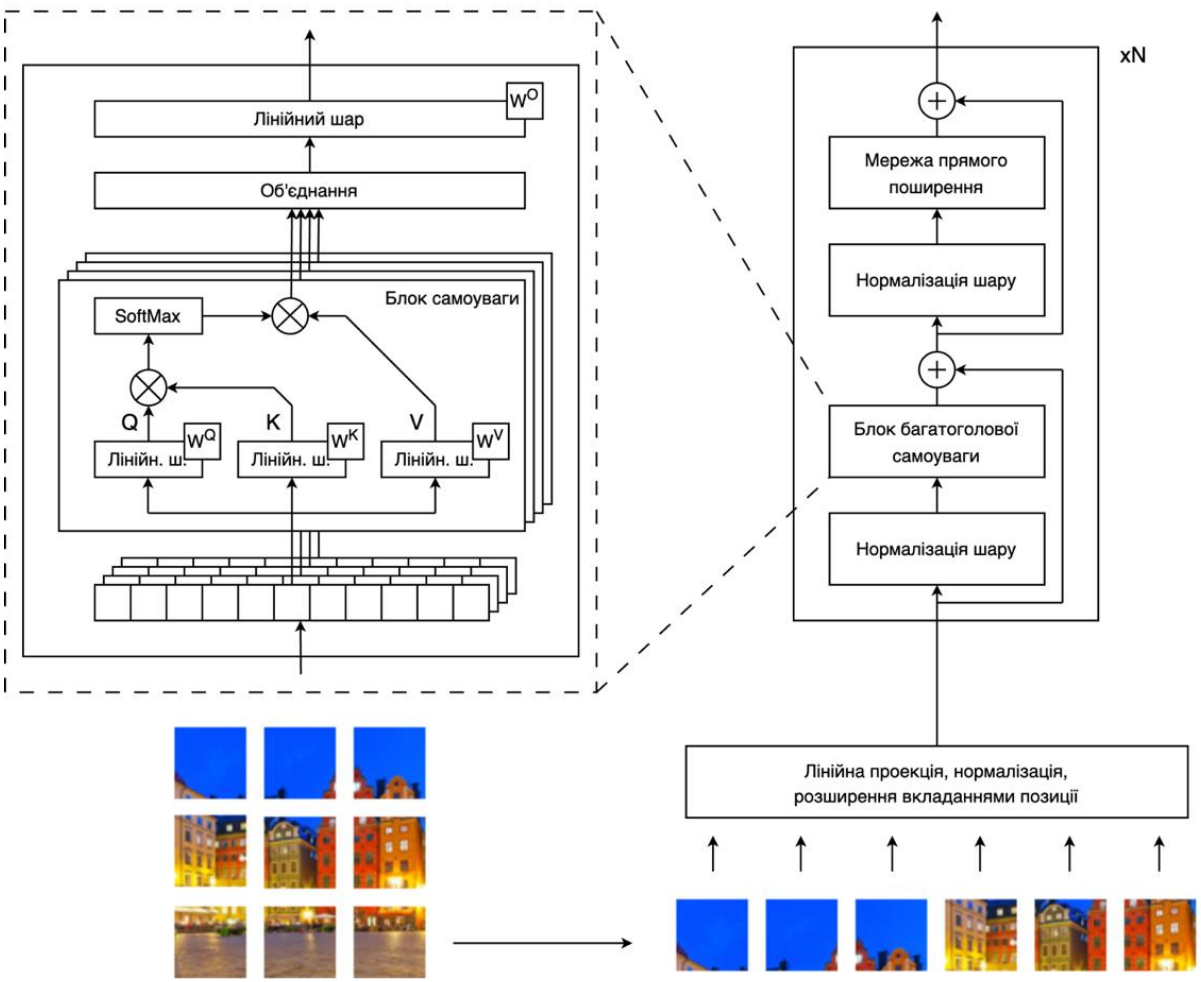

Рис. 2. Загальна схема Visual Transformer

Архітектура трансформер розроблена для обробки послідовностей, проте вона не враховує явно позицію кожного фрагмента у послідовності. Щоб усунути це обмеження застосовують вкладення позиції, що кодують позицію кожного фрагмента на зображенні. Вкладення фрагментів та відповідні їм вкладення позиції об'єднують перед передачею на вхід до блоків трансформера. Цей механізм дозволяє моделі врахувати відносну позицію фрагментів та виділити просторову інформацію з зображення.

Ядром архітектури трансформер є механізм самоуваги, який моделює взаємодії і зв'язки між фрагментами у вхідній послідовності. Результат роботи функції самоуваги можна означити як зважену суму вхідних значень, де вага надана кожному значенню (вага уваги) визначається функцією сумісності запиту і відповідного ключа. Розглянемо послідовність з n вкладень $\{X_1, X_2, X_3, \dots X_n\}$, де $X \in \mathbb{R}^{n \times d}$, та d - розмір вкладення, і визначені матриці ваг, що навчаються $W^Q \in \mathbb{R}^{n \times d_Q}$, $W^K \in \mathbb{R}^{n \times d_K}$, $W^V \in \mathbb{R}^{n \times d_V}$ для лінійних проекцій запитів, ключів і значень відповідно, тоді самоувагу можна визначити так:

$$Q = X \cdot W^Q, \#(10)$$
$$K = X \cdot W^K, \#(11)$$
$$V = X \cdot W^V, \#(12)$$
$$\text{Attention}(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_Q}}\right)V. \#(13)$$

У [19] показано, що застосування механізму самоуваги декілька разів паралельно до однієї і тієї ж послідовності надає моделі можливість "зосереджуватись" на інформації з різних підпросторів представлення для різних комбінації фрагментів у вхідній послідовності. Для

чого самоувагу рахують h разів, проєктуючи вхідну послідовність X за допомогою окремих наборів ваг W_i^Q, W_i^K, W_i^V. Кожен застосований таким чином механізм самоуваги називають головою самоуваги, їх результати об'єднуються і проєктуються за допомогою матриці ваг W^O:

$$head = \text{Attention}\left(XW_i^Q, XW_i^K, XW_i^V\right), \#(14)$$
$$MultiHead(X) = \text{Concat}(head_1, \dots\ head_2)W^O, \#(15)$$

де $W_i^Q \in \mathbb{R}^{n\times d_Q}$, $W_i^K \in \mathbb{R}^{n\times d_K}$, $W_i^V \in \mathbb{R}^{n\times d_V}$, $W^O \in \mathbb{R}^{hd_v\times n}$. З метою зменшення обчислювального навантаження при обчислені багатоголової самоуваги, кожна голова працює лише з частиною кожного вкладення, тобто: $d_Q = d_k = d_v = \frac{d}{h}$.

Основною перевагою механізму самоуваги порівняно з механізмом згортки є те, що вага уваги обчислюється динамічно в залежності від вхідних значень, на відміну від ваг фільтрів, які є статичними для всього вхідного набору даних. У [21] показано, що самоувага, за умов достатньої кількості параметрів, є напрочуд гнучким процесом, і дозволяє успішно виділяти як глобальні так і локальні ознаки. Слід також зазначити, що застосування архітектури трансформер вимагає істотно більших наборів тренувальних даних, через значно більшу ємність мережі.

Застосування архітектури трансформер до задачі super-resolution

Архітектуру трансформер було вперше застосовано до задачі SR у роботі [22]. Запропонована мережа отримала назву **IPT** (Image processing transformer) і складається з вхідного компоненту для виділення ознак з вхідного зображення, тіла та вихідного компонента для відновлення зображення з набору ознак. Вхідний і вихідний компоненти відрізняються в залежності від типу задачі: знешумлення, SR або видалення дощу. Тіло складається з 12-ти блоків кодування і 12-ти блоків декодування, кожен з яких побудований аналогічно до блоків ViT. Вихідний компонент складається зі згорткового шару і двох шарів ResNet. Вхідні ознаки розбиваються на фрагменти і представляються у вигляді вкладень з позиційним кодуванням аналогічно до ViT. У випадку SR вихідний компонент складається з одного або двох субпіксельних згорткових шарів [12]. Запропонована модель показала приріст продуктивності для коефіцієнтів збільшення x2, x3, x4 для всіх наборів даних у задачі SR порівняно з актуальними на той час ЗНМ, наприклад RCAN. Проте, необхідно зазначити, що модель IPT містить 114M параметрів проти 16M у RCAN. Окрім того, було показано, що при навчанні на обмеженому наборі даних (менше ніж 60 % набору даних ImageNet [7]) IPT показує гіршу продуктивність ніж актуальні ЗНМ, проте продуктивність зростає при збільшенні розмірів тренувального набору.

Ефективне застосування архітектури трансформер до задач комп'ютерного зору пов'язане з викликами, які витікають з відмінностей між візуальним і мовленнєвим доменами. Перша відмінність – масштаб. Зазвичай зображення містять візуальні елементи різних масштабів, що ускладнює їх обробку за допомогою архітектури трансформер, який аналогічно до обробки токенів в ОПМ, працює з фрагментами одного розміру. Друга – об'єм інформації, адже обчислювальна складність обрахунку самоуваги квадратична довжині вхідної послідовності, що стає особливо критичним під час обробки зображень високої роздільної здатності.

У [23] представлено Swin Transformer – візуальний трансформер загального призначення, який пропонує підходи до вирішення означених складнощів. Для підвищення ефективності пропонується застосувати механізм локальної самоуваги, при якому самоувага буде обраховуватися не для всього набору вхідних вкладень, а лише для його частини $N \times N$ фрагментів. Це дозволить отримати лінійну обчислювальну складність відносно розмірів зображення. При цьому, для збереження зв'язків між візуальними елементами які потрапили до різних вікон необхідно "зсувати" вікна при обрахунку самоуваги в глибших шарах мережі.

Мережа **SwinIR** [24] заснована на ідеях з Swin Transformer, показала приріст PSNR у межах 0.08-0.28dB відносно **IPT**, за умови значно менших розмірів 11.8M і навчанні на значно меншому наборі даних, заклавши ґрунтовну основу для майбутніх досліджень. **SwinIR** побудована за архітектурою подібною до RCAN, що показано на рис. 3 і складається з: модуля виділення поверхневих ознак, модуля виділення глибоких ознак і модуля відновлення зображення високої роздільної здатності. Модуль виділення поверхневих ознак являє собою згортковий шар з ядром 3x3, і забезпечує виділення поверхневих ознак та переведення зображення в простір вищої мірності для подальшої обробки модулем виділення глибоких ознак. Модуль виділення глибоких ознак складається з N_{RG} RG (residual group – залишкових груп) та згорткового шару. Кожна RG складається з N_{TB} TB (transformer block – блок трансформера) і згорткового шару. У випадку **SwinIR** TB – блок трансформера ViT (рис. 2), з тією відмінністю, що тут застосована локальна самоувага з зсувними вікнами. Поверхневі і глибокі ознаки агрегуються перед модулем відновлення зображення високої роздільної здатності, який у випадку задачі SR являє собою субпіксельний згортковий шар [12].

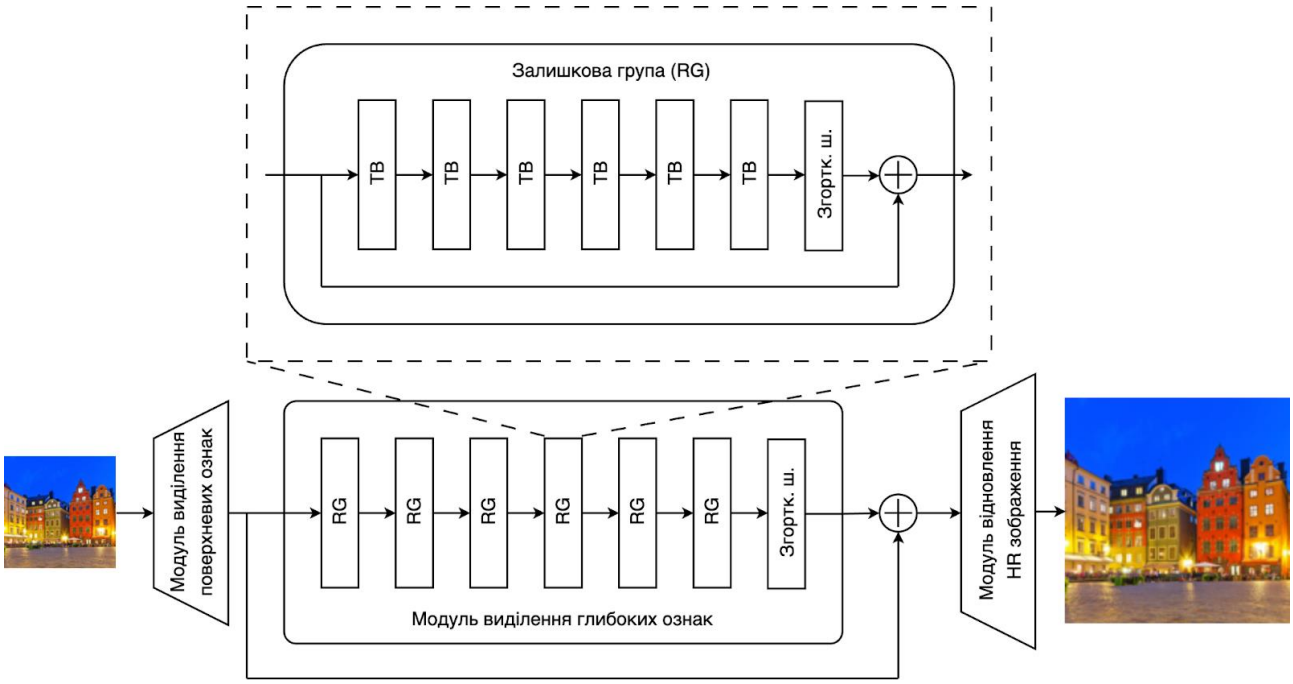

Рис. 3. Архітектура мережі SwinIR

Мережі запропоновані в подальших роботах мають архітектуру подібну до **SwinIR** і зосереджені, головним чином, на пошуках ефективного способу залучення більшої кількості глобальної інформації за умови збереження розмірів мережі, розмірів вікон локальної самоуваги та розміру тренувального набору даних. У мережі **EDT** (encode-decoder-based transformer) [25] запропонований підхід, при якому вхідну мапу ознак ділять на дві рівні частини по канальній розмірності і до кожної половини застосовують прямокутні вікна самоуваги вертикальної або горизонтальної орієнтації, таким чином утворюючи хрестоподібне рецептивне поле. У [26] запропоновано **ART** (attention retractable transformer), тут парні блоки у RG замінені на блоки SAB (sparse attention block – блок розрідженої самоуваги), у яких самоувага застосовується до фрагментів, які розташовані через певний інтервал один від одного. Подібний підхід запропоновано у **DWT** (Detailed window transformer) [27], проте тут інтервал між фрагментами зростає із поглибленням мережі. У [28] запропоновано RWin-SA (Rectangle-window self-attention) – TB з прямокутними вікнами самоуваги що перетинаються, подібно до **EDT**, проте вікна різної орієнтації застосовуються до різних голів самоуваги. Досліджено прямокутні вікна самоуваги – мережа CAT-R, та прямокутні вікна для яких одна сторона дорівнює довжині чи ширині зображення - мережа

CAT-A. Окрім того, RWin-SA розширено за допомогою модуля LCM (Locality complementary module), який являє собою згортковий шар по V, розташований паралельно блоку самоуваги. Застосування ковзних вікон самоуваги досліджено у **Uniwin** [29].

У [30] запропоновано **SRFormer**, що складається з блоків PSA (Permuted self-attention), які шляхом зменшення канальної розмірності K та V дозволяють збільшити ефективність обрахунку самоуваги, і як результат, збільшити розміри вікна самоуваги із збереженням кількості параметрів мережі і обчислювальної складності.

У трансформері **SwinFIR** [31] дослідили можливість застосування частотного представлення інформації замінивши згортковий шар у кожній RG на SFB (spatial-frequency block), який складається з двох гілок – частотної і просторової. Частотна гілка аналогічна запропонованій у [32] і виконує послідовно згортки прямого та зворотного перетворення Фур'є з метою виділення глобальних ознак. Просторова складається з двох послідовних згорткових шарів.

Застосування канальної самоуваги досліджено в [33]. Тут запропонований **HAT** (hybrid attention transformer), де в ТВ паралельно до блоку самоуваги додано CAB (channel attention block – блок канальної уваги) аналогічний до блоку канальної уваги у RCAN. Також у HAT замінили останній згортковий шар у кожній RG на OCAB (overlapping cross-attention block). На відміну від самоуваги в звичайному ТВ, де Q, K, V рахуються для вікон однакового розміру, в OCAB K та V рахуються для вікна, більшого за те, для якого рахується Q, що має сприяти утворенню міжвіконних зв'язків. У [34] запропоновано мережу **DAT** (dual aggregation transformer), у якій блоки просторової DSTB (dual spatial transformer block) та канальної уваги DCTB (dual channel transformer block) застосовуються почергово в межах RG. Окрім того, кожен ТВ у цій мережі розширено згортковим шаром по V паралельно до блоку уваги і модулем AIM (adaptive interaction module), який дозволяє ефективно об'єднати ознаки, отримані з блоку самоуваги та згорткового шару. Такий підхід дозволяє ефективно об'єднувати ознаки канальної і просторової розмірностей як на рівні ТВ так і на рівні модуля виділення глибоких ознак, що має позитивно вплинути на репрезентативну можливість мережі.

У [35] досліджено можливість попередньої агрегації глобальної інформації перед обчисленням локальної самоуваги. Для цього запропоновано RGM (recursive generalisation module), який за допомогою рекурсивного застосування згорткового шару до вхідної карти ознак дозволяє отримати стиснуту карту ознак. І блок RA-SA (recursive-generalization self-attention), який базується на Rwin-SA, містить RGM, і обчислює значення K та V на основі стисненої карти ознак, але Q на основі відповідного вікна локальної самоуваги. На основі блоку RA-SA побудовано мережу **RGT** (recursive generalization transformer), у якій блоки RA-SA послідовно чергуються з блоками Rwin-SA.

У таблиці 1 наведено порівняння характеристик і продуктивності розглянутих вище мереж на основі тестового набору даних Urban100 [36] з коефіцієнтом збільшення x4. Для порівняння наведено актуальні ЗНМ з механізмами канальної та нелокальної розрідженої уваги – RCAN та NLSA [37]. За основу для порівняння обрано мережу **SwinIR**, у колонках ΔPSNR та ΔSSIM наведено зміну зазначених метрик відносно **SwinIR**.

Очевидно, що площа вікна уваги безпосередньо впливає на продуктивність, що підтверджує важливість глобальної інформації для вирішення задачі SR. Таким чином пошук ефективних шляхів залучення якомога більшої кількості глобальної інформації залишається актуальним. Наразі, найкращі результати показали **DWT**, де застосована розріджена самоувага зі змінним інтервалом та **Uniwin,** у якому запропоновано застосування ковзних вікон уваги. Слід також відзначити підхід, застосований у **RGT**, де пропонується використати рекурсивний згортковий шар для стиснення вхідної мапи ознак по просторовій розмірності перед обчисленням самоуваги.

Таблиця 1

Порівняння параметрів і продуктивності мереж SR, побудованих на основі архітектури трансформер

Дата публікації	Мережа	Навчальний набір	Розмір вікна самоуваги	К-сть параметрів $\times 10^6$	Urban100 (x4)			
					PSNR	SSIM	ΔPSNR	ΔSSIM
2018	**RCAN**	DIV2K		16.0	26.82	0.8087	-0.63	-0.0167
12.2020	**IPT**	ImageNet		115.5	27.26		-0.19	
2021	**NLSA**	DIV2K [11]			26.96	0.8109	-0.49	-0.0145
08.2021	**SwinIR**		8x8	11.8	27.45	0.8254	0	0
12.2021	**EDT**	DF2K[11, 38]	6x24	11.7	27.46	0.8246	0.01	-0.0008
05.2022	**HAT**	DF2K	16x16	20.8	27.97	0.8368	**0.52**	**0.0114**
08.2022	**SwinFIR**	DF2K	12x12	14.0	27.87	0.8348	0.42	0.0094
01.2022	**ART**	DF2K	8x8	16.5	27.77	0.8321	0.32	0.0067
11.2022	**CAT-R**	DF2K	4x16	16.6	27.62	0.8292	0.17	0.0038
11.2022	**CAT-A**	DF2K	4xW[H]	16.6	27.89	0.8339	0.44	0.0085
03.2023	**RGT**	DF2K	8x32	13.3	27.98	0.8369	**0.53**	**0.0115**
03.2023	**SRFormer**	DF2K	24x24	10.4	27.68	0.8311	0.23	0.0057
05.2023	**DWT**	DF2K	16x16	12.0	27.81	0.8324	0.36	0.0070
08.2023	**DAT**	DF2K	8x32	14.8	27.87	0.8343	0.42	0.0089
02.2024	**Uniwin**	DF2K	9x9	12.0	27.90	0.8362	0.45	0.0108

Розширення блоку трансформера згортковими шарами паралельно до блоків самоуваги, що запропоновано у мережах **CAT**, **HAT** та **DAT** також складає позитивний ефект на продуктивність у задачі SR. Це може свідчити або про обмежені можливості трансформера у виділенні локальних ознак, або нестачу просторової інформації і потребує подальшого дослідження.

Мережі **HAT** та **DAT** також показали високу продуктивність, що свідчить про позитивний ефект від застосування канальної уваги. Отже, ознаки в канальній розмірності також мають різну вагу, що робить скорочення канальної розмірності цікавим напрямком досліджень, адже це має дозволити не лише збільшити продуктивність, а і скоротити час обчислення. Подібний підхід застосовано у **SRFormer**.

Варто відзначити високе значення метрики SSIM за умови невеликого вікна самоуваги для мережі **SwinFIR**, що може свідчити про позитивний ефект від застосування частотного представлення інформації для задачі SR і також вимагає подальшого дослідження.

Ключовим аспектом механізму самоуваги у архітектурі трансформер є можливість зосереджуватись на важливій інформації з потоку даних, що також є невід'ємною властивістю біологічної системи людини [39]. Тож перспективною є реалізація механізму самоуваги на основі спайкінгових нейронних мереж (СНМ) [40, 41. У [42] запропоновано поєднання архітектури трансформер та СНМ [43] для вирішення задачі класифікації зображень. Дослідити аналогічний підхід доречно і у випадку задачі SR. Застосування СНМ [44, 45] дозволить забезпечити вищий рівень енергоефективності, а також надасть можливість ефективного вирішення задачі в умовах реального часу.

Висновки

1. Застосування архітектури трансформер до задачі SR дозволило досягти значного приросту продуктивності (ΔPSNR: 0.5-1.2 dB, ΔSSIM: 0.0055-0.0234) порівняно з актуальними підходами на основі глибоких нейронних мереж, таких як ЗНМ або ГЗНМ.

2. Проте, застосування архітектури трансформер до задачі SR пов'язане з рядом викликів, а саме: висока обчислювальна складність у випадку застосування глобальної самоуваги, обмеженість у захопленні просторової інформації, необхідність пошуку компромісу між обчислювальною складністю і об'ємом залученої глобальної інформації,

висока ємність мереж, заснованих на архітектурі трансформер і, як наслідок, необхідність у великих об'ємах тренувальних даних.

3. Проаналізовані роботи, у більшості своїй, сфокусовані на пошуку компромісу між об'ємом залученої глобальної інформації і обчислювальною складністю. Для чого впроваджують і досліджують різні форми локальної самоуваги, і, наразі, можна стверджувати, що розріджена самоувага забезпечує найкращий результат. Альтернативним підходом є запропонований у RGT метод стиснення вхідної мапи ознак перед застосуванням самоуваги. Комбінація архітектури трансформер зі ЗНМ, застосування канальної самоуваги та використання частотного представлення інформації також є перспективними напрямками досліджень.

4. З метою ефективного вирішення задачі SR в умовах реального часу є доречним дослідити можливість реалізації механізму самоуваги з використанням СНМ.

СПИСОК ЛІТЕРАТУРИ

1. New edge-directed interpolation [Electronic resource] / Li Xin, M. T. Orchard // IEEE Transactions on Image Processing. – 2001. – Vol. 10, № 10. – P. 1521 – 1527. – Access mode : https://doi.org/10.1109/83.951537 (date of access: 15.02.2024).

2. SoftCuts: A Soft Edge Smoothness Prior for Color Image Super-Resolution [Electronic resource] / Shengyang Dai, Mei Han, Wei Xu [et al.] // IEEE Transactions on Image Processing. – 2009. – Vol. 18, № 5. – P. 969 – 981. – Access mode : https://doi.org/10.1109/tip.2009.2012908 (date of access: 15.02.2024).

3. Image super-resolution using gradient profile prior [Electronic resource] / Jian Sun, Zongben Xu, Heung-Yeung Shum // 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, 23–28 June 2008. – Access mode : https://doi.org/10.1109/cvpr.2008.4587659 (date of access: 15.02.2024).

4. Super-resolution through neighbor embedding [Electronic resource] / Hong Chang, Dit-Yan Yeung, Yimin Xiong // Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA. – Access mode : https://doi.org/10.1109/cvpr.2004.1315043 (date of access: 15.02.2024).

5. Image Deblurring and Super-Resolution by Adaptive Sparse Domain Selection and Adaptive Regularization [Electronic resource] / Weisheng Dong [et al.] // IEEE Transactions on Image Processing. – 2011. – Vol. 20, № 7. – P. 1838 –1857. – Access mode : https://doi.org/10.1109/tip.2011.2108306 (date of access: 15.02.2024).

6. Fast image super resolution via local regression [Electronic resource] / Gu Shuhang, Sang Nong, Ma Fan // Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, 11–15 October 2012. – P. 3128 – 3131. – Access mode : https://ieeexplore.ieee.org/document/6460827 (date of access: 15.02.2024).

7. ImageNet classification with deep convolutional neural networks [Electronic resource] / Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton // Communications of the ACM. – 2017. – Vol. 60, № 6. – P. 84 – 90. – Access mode : https://doi.org/10.1145/3065386 (date of access: 15.02.2024).

8. Learning a Deep Convolutional Network for Image Super-Resolution [Electronic resource] / Dong Chao, Chen Change Loy, Kaiming He [et al.] // Computer Vision – ECCV 2014, 6–12 September 2014. – P. 184 – 199. – Access mode : https://doi.org/10.1007/978-3-319-10593-2_13 (date of access: 15.02.2024).

9. Accurate Image Super-Resolution Using Very Deep Convolutional Networks [Electronic resource] / Jiwon Kim, Jung Kwon Lee, Kyoung Mu Lee // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. – Access mode : https://doi.org/10.1109/cvpr.2016.182 (date of access: 15.02.2024).

10. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [Electronic resource] / Christian Ledig, Lucas Theis; Ferenc Huszár [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017. – Access mode : https://doi.org/10.1109/cvpr.2017.19 (date of access: 15.02.2024).

11. Enhanced Deep Residual Networks for Single Image Super-Resolution [Electronic resource] / Bee Lim, Sanghyun Son, Heewon Kim [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017. – Access mode : https://doi.org/10.1109/cvprw.2017.151 (date of access: 15.02.2024).

12. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network [Electronic resource] / Wenzhe Shi, Jose Caballero, Ferenc Huszár [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. – Access mode : https://doi.org/10.1109/cvpr.2016.207 (date of access: 15.02.2024).

13. Image Super-Resolution Using Very Deep Residual Channel Attention Networks [Electronic resource] / Zhang Yulun, Kunpeng Li, Kai Li [et al.] // Computer Vision – ECCV 2018, Munich, 8–14 September 2018. – P. 294 – 310. – Access mode : https://doi.org/10.1007/978-3-030-01234-2_18 (date of access: 15.02.2024).

14. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks [Electronic resource] / Wang Xintao, Ke Yu, Shixiang Wu [et al.] // Computer Vision - ECCV 2018 Workshops, Munich, 8–14 September 2018. – P. 63 – 79. – Access mode : https://doi.org/10.1007/978-3-030-11021-5_5 (date of access: 15.02.2024).

15. SRDiff: Single image super-resolution with diffusion probabilistic models [Electronic resource] / Haoying Li, Yifan Yang, Meng Chang [et al.] // Neurocomputing. – 2022. – Vol. 479. – P. 47 – 59. – Access mode : https://doi.org/10.1016/j.neucom.2022.01.029 (date of access: 15.02.2024).

16. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution [Electronic resource] / Wei-Sheng Lai Jia-Bin Huang; Narendra Ahuja [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017. – 2017. – Access mode : https://doi.org/10.1109/cvpr.2017.618 (date of access: 15.02.2024).

17. Image Quality Assessment: From Error Visibility to Structural Similarity [Electronic resource] / Z. Wang A. C. Bovik, H. R. Sheikh [et al.] // IEEE Transactions on Image Processing. – 2004. – Vol. 13, № 4. – P. 600 – 612. – Access mode : https://doi.org/10.1109/tip.2003.819861 (date of access: 15.02.2024).

18. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [Electronic resource] / A. Dosovitskiy, L. Beyer, A. Kolesnikov [et al.] // International Conference on Learning Representations, 3–7 May 2021.– Access mode : https://openreview.net/pdf?id=YicbFdNTTy (date of access: 15.02.2024).

19. Attention is All you Need [Electronic resource] / Ashish Vaswani, Noam Shazeer, Niki Parmar [et al.] // Advances in Neural Information Processing Systems, 4 – 9 December 2024. – Access mode : https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (date of access: 15.02.2024).

20. Early Convolutions Help Transformers See Better [Electronic resource] / Xiao Tete, Mannat Singh, Eric Mintun [et al.] // Advances in Neural Information Processing Systems: 2021, 6 – 14 December 2021. – Access mode : https://proceedings.neurips.cc/paper/2021/hash/ff1418e8cc993fe8abcfe3ce2003e5c5-Abstract.html (date of access: 15.02.2024).

21. On the Relationship between Self-Attention and Convolutional Layers [Electronic resource] / Cordonnier Jean-Baptiste, Loukas Andreas, Martin Jaggi // International Conference on Learning Representations , 27 – 30 April 2020. – Access mode : https://openreview.net/forum?id=HJlnC1rKPB (date of access: 15.02.2024).

22. Pre-Trained Image Processing Transformer [Electronic resource] / Hanting Chen, Yunhe Wang, Tianyu Guo [et al.] // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20 – 25 June 2021. – Access mode : https://doi.org/10.1109/cvpr46437.2021.01212 (date of access: 15.02.2024).

23. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows [Electronic resource] / Ze Liu, Yutong Lin, Yue Cao, Han Hu [et al.] // 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. – Access mode : https://doi.org/10.1109/iccv48922.2021.00986 (date of access: 15.02.2024).

24. SwinIR: Image Restoration Using Swin Transformer [Electronic resource] / Jingyun Liang, Jiezhang Cao, Guolei Sun [et al.] // 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October. 2021. – Access mode : https://doi.org/10.1109/iccvw54120.2021.00210 (date of access: 15.02.2024).

25. On Efficient Transformer-Based Image Pre-training for Low-Level Vision [Electronic resource] / Wenbo Li, Xin Lu, Shengju Qian, [et al.] // International Joint Conference on Artificial Intelligence, Macao, 19–25 August 2024. – Access mode : https://www.ijcai.org/proceedings/2023/0121.pdf (date of access: 15.02.2024).

26. Accurate Image Restoration with Attention Retractable Transformer [Electronic resource] / Jiale Zhang, Yulun Zhang, Jinjin Gu [et al.] // The Eleventh International Conference on Learning Representations, Kigali, 30 April – 5 May 2023. – Access mode : https://openreview.net/pdf?id=IloMJ5rqfnt (date of access: 15.02.2024).

27. Image Super-Resolution Using Dilated Window Transformer [Electronic resource] / Soobin Park, Yong Suk Choi // IEEE Access. – 2023. – P. 1. – Access mode : https://doi.org/10.1109/access.2023.3284539 (date of access: 15.02.2024).

28. Cross Aggregation Transformer for Image Restoration [Electronic resource] /Zheng Chen, Yulun Zhang, Jinjin Gu [et al.] // Advances in Neural Information Processing Systems, New Orleans, 11–19 December 2022. – Access mode: https://openreview.net/forum?id=wQ2QNNP8GtM (date of access: 15.02.2024).

29. Image Super-Resolution with Unified-Window Attention [Electronic resource] / Gunhee Cho, Yong Suk Choi // IEEE Access. – 2024. – P. 1. – Access mode : https://doi.org/10.1109/access.2024.3368436 (date of access: 15.02.2024).

30. SRFormer: Permuted Self-Attention for Single Image Super-Resolution [Electronic resource] / Yupeng Zhou, Zhen Li, Chun-Le Guo [et al.] // 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. – Access mode : https://doi.org/10.1109/iccv51070.2023.01174 (date of access: 15.02.2024).

31. SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution [Electronic resource] / Dafeng Zhang, Feiyu Huang, Shizhuo Liu [et al.]. : arxiv.org, 2023. – 14 p. – Access mode : https://arxiv.org/pdf/2208.11247.pdf (date of access: 15.02.2024).

32. Resolution-robust Large Mask Inpainting with Fourier Convolutions [Electronic resource] / Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin [et al.] // 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022. – Access mode : https://doi.org/10.1109/wacv51458.2022.00323 (date of access: 15.02.2024).

33. Activating More Pixels in Image Super-Resolution Transformer [Electronic resource] / Xiangyu Chen, Xintao Wang, Jiantao Zhou [et al.] // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023. – Access mode : https://doi.org/10.1109/cvpr52729.2023.02142 (date of access: 15.02.2024).

34. Dual Aggregation Transformer for Image Super-Resolution [Electronic resource] / Zheng Chen, Yulun Zhang, Jinjin Gu [et al.] // 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. – Access mode : https://doi.org/10.1109/iccv51070.2023.01131 (date of access: 15.02.2024).

35. Recursive Generalization Transformer for Image Super-Resolution [Electronic resource] / Zheng Chen, Yulun Zhang, Jinjin Gu [et al.] // The Twelfth International Conference on Learning Representations, Vienna, 7–11 May 2024. – Access mode : https://openreview.net/forum?id=owziuM1nsR (date of access: 15.02.2024).

36. Single image super-resolution from transformed self-exemplars [Electronic resource] / Jia-Bin Huang, Abhishek Singh, Narendra Ahuja // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. – Access mode : https://doi.org/10.1109/cvpr.2015.7299156 (date of access: 15.02.2024).

37. Image Super-Resolution with Non-Local Sparse Attention [Electronic resource] / Yiqun Mei, Yuchen Fan, Yuqian Zhou // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20 – 25 June 2021. – Access mode : https://doi.org/10.1109/cvpr46437.2021.00352 (date of access: 15.02.2024).

38. NTIRE 2017 Challenge on Single Image Super-Resolution: Methods and Results [Electronic resource] / Radu Timofte, Eirikur Agustsson, Luc Van Gool [et al.] // 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21 – 26 July 2017. – Access mode : https://doi.org/10.1109/cvprw.2017.149 (date of access: 15.02.2024).

39. Relating transformers to models and neural representations of the hippocampal formation [Electronic resource] / James C. R. Whittington, Joseph Warren, Tim E. J Behrens // International Conference on Learning Representations, 25–29 April 2022. – Access mode : https://openreview.net/forum?id=B8DVo9B1YE0 (date of access: 15.02.2024).

40. Бардаченко В. Ф. Перспективи застосування імпульсних нейронних мереж з таймерним представленням інформації для розпізнавання динамічних образів / В. Ф .Бардаченко, О. К. Колесницький, С. А. Василецький // УСіМ. – 2003. – № 6. - С. 73 – 82.

41. Колесницький О. К. Принципи побудови архітектури спайкових нейрокомп'ютерів / О. К. Колесницький // Вісник Вінницького політехнічного інституту. – Вінниця: УНІВЕРСУМ-Вінниця. – 2014. – № 4 (115). – С. 70 – 78.

42. Spikformer: When Spiking Neural Network Meets Transformer [Electronic resource] / Zhaokun Zhou, Yuesheng Zhu, Chao He [et al.] // The Eleventh International Conference on Learning Representations, Kigali, 1–5 May 2023. – Access mode : https://openreview.net/forum?id=frE4fUwz_h (date of access: 15.02.2024).

43. Optoelectronic implementation of pulsed neurons and neural networks using bispin-devices [Electronic resource] / O. K. Kolesnytskyj, I. V. Bokotsey, S. S. Yaremchuk // Optical Memory and Neural Networks. – 2010. – Vol. 19, № 2. – P. 154 – 165. – Access mode : https://doi.org/10.3103/s1060992x10020062 (date of access: 15.02.2024).

44. Optoelectronic spiking neural network [Electronic resource] / V. P. Kozemiako, O. K. Kolesnytskyj, T. S. Lischenko [et al.] // Optical Fibers and Their Applications 2012, Krasnobrod, Poland. – 2013. – Access mode : https://doi.org/10.1117/12.2019340 (date of access: 25.04.2024).

45. Neurocomputer architecture based on spiking neural network and its optoelectronic implementation [Electronic resource] / Oleh K. Kolesnytskyj, Vladislav V. Kutsman, Krzysztof Skorupski [et al.] // Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019, Wilga, Poland, 25 May – 2 June 2019 / ed. by R. S. Romaniuk, M. Linczuk. – [S. l.], 2019. – Access mode : https://doi.org/10.1117/12.2536607 (date of access: 25.04.2024).

Стаття надійшла до редакції 22.02.2024.
Стаття пройшла рецензування 25.02.2024.

Козлов Сергій Леонідович – аспірант кафедри комп'ютерних наук, e-mail: serhii.kozlov@gmail.com.
Колесницький Олег Костянтинович – канд. тех. наук, професор кафедри комп'ютерних наук.
Вінницький національний технічний університет.