

ЗНЕСУМЛЕННЯ ЗОБРАЖЕНЬ НА ОСНОВІ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Задорожний М. О., студент 2 курсу магістратури, групи ІІІ

Анотація. Дана робота присвячена дослідженню та практичній реалізації задачі знесумлення цифрових зображень із використанням сучасних методів глибокого навчання. Знесумлення є однією з ключових задач комп'ютерного зору, яка полягає у відновленні чистого зображення з його зашумленої версії без істотної втрати деталей та структури. У межах роботи розглянуто застосування нейронної мережі архітектури NAFNet (Nonlinear Activation Free Network), орієнтованої на задачі відновлення зображень, а також використання модифікованого датасету на основі COCO 2017 для формування навчальних прикладів. Проведено аналіз особливостей архітектури моделі, її структурних блоків, принципу формування шуму та підготовки даних.

Ключові слова: знесумлення зображень, *image denoising*, глибоке навчання, *NAFNet*, *COCO*, реконструкція зображень, комп'ютерний зір

Abstract. The work is dedicated to the study and practical implementation of the image denoising problem using modern deep learning techniques. Image denoising is a fundamental task in computer vision aimed at restoring clean images from their noisy versions while preserving spatial details and structural integrity. Within this research, the NAFNet (Nonlinear Activation Free Network) architecture is applied, which is specifically designed for low-level image restoration tasks, alongside a modified dataset based on COCO 2017. The paper analyzes the dataset structure, preprocessing pipeline, noise modeling process, and the internal architecture of the proposed neural network, emphasizing its suitability for high-quality denoising.

Keywords: *image denoising, deep learning, NAFNet, COCO dataset, image restoration, convolutional neural network.*

1. Вступ

Постановка проблеми та актуальність

Знесумлення цифрових зображень є однією з фундаментальних задач низькорівневої обробки в комп'ютерному зорі, яка безпосередньо впливає на якість і надійність усіх подальших етапів аналізу візуальної інформації. Практично будь-яка система комп'ютерного зору — від систем відеоспостереження та медичної діагностики до автономних транспортних засобів і мобільних застосунків — потребує високоякісного вхідного зображення. Присутність шумів призводить до деградації текстур, зменшення контрасту, спотворення контурів об'єктів і, як наслідок, до зниження точності алгоритмів детекції, класифікації та сегментації.

Шум у цифрових зображеннях виникає через сукупність фізичних та технічних факторів, серед яких: обмежена чутливість сенсорів, теплові перешкоди, електронні флуктуації, нестабільність освітлення, високе значення ISO, втрати при стисненні та передачі даних. У реальних умовах шум має складну природу і часто поєднує кілька моделей, таких як імпульсний та структурний шум, що ускладнює задачу його усунення.

Класичні алгоритми знесумлення базуються на математичних та статистичних методах, серед яких: гаусівська фільтрація, медіанні фільтри, *bilateral filtering*, вейвлет-перетворення, методи на основі частотного аналізу та варіаційні підходи. Хоча ці методи є обчислювально дешевими, їх застосування супроводжується істотним компромісом між пригніченням шуму та збереженням структурних деталей. Як правило, агресивне згладжування призводить до розмиття меж і втрати дрібних елементів, тоді як м'яка фільтрація не забезпечує достатнього очищення зображення.

Із розвитком глибокого навчання відбувається зміщення фокусу від ручного проєктування фільтрів до навчання моделей, здатних апроксимувати складне перетворення "зашумлене →

чисте зображення". Перші успішні спроби застосування згорткових нейронних мереж для цієї задачі представлені моделями DnCNN[1], RED-Net[2], FFDNet[3], IRCNN[4], які продемонстрували значне покращення якості результуючих зображень.

Однак сучасні тенденції розвитку знешумлення спрямовані на використання надглибоких архітектур, зокрема transformer-based моделей (SwinIR[5], Restormer[6], Uformer[7]) та дифузійних моделей (Diffusion Models[8], DDPM[8]). Такі системи забезпечують рекордні показники якості, проте їх характерною рисою є надзвичайно велика кількість параметрів — від десятків до сотень мільйонів, що обумовлює:

- високі вимоги до GPU-ресурсів;
- збільшення часу навчання;
- високе енергоспоживання;
- неможливість застосування в режимі реального часу на малопотужних пристроях;
- складність масштабування.

На цьому фоні особливої значущості набувають компактні й оптимізовані архітектури, здатні забезпечувати якісне знешумлення без надмірної параметризації. Однією з таких моделей є NAFNet (Nonlinear Activation Free Network)[9], яка демонструє конкурентну якість при значно меншій кількості параметрів. Типові конфігурації NAFNet мають у діапазоні 2–16 млн параметрів, тоді як сучасні state-of-the-art моделі можуть перевищувати 100–300 млн, що робить NAFNet більш придатною для практичного застосування в реальних системах з обмеженими ресурсами.

Таким чином, актуальність теми полягає у необхідності вирішення протиріччя між зростаючими вимогами до якості відновлення зображень та обмеженнями апаратного забезпечення. Розробка і дослідження ефективних моделей, таких як NAFNet, дозволяє забезпечити оптимальний баланс між якістю, швидкістю та ресурсною ефективністю.

Метою даної роботи є дослідження підходу до знешумлення зображень із використанням архітектури NAFNet та аналіз її ефективності у порівнянні з сучасними надпараметризованими моделями глибокого навчання.

Аналіз існуючих підходів до знешумлення зображень

У сучасній літературі задачу знешумлення зображень прийнято класифікувати за типом архітектурних рішень та принципом моделювання процесу деградації.

Методи традиційної обробки: Включають лінійні та нелінійні фільтри, варіаційні моделі, методи на основі регуляризації, вейвлет-декомпозицію. Вони прості у реалізації, але демонструють обмежені результати для складних типів шуму.

CNN-архітектури: Моделі типу DnCNN[1], U-Net[10], FFDNet[3] будуються на основі залишкових зв'язків і дозволяють ефективно навчатися на парах зображень. Їх кількість параметрів знаходиться в межах 1–10 млн, що забезпечує помірний баланс між якістю та продуктивністю.

Transformer-based підходи: Моделі SwinIR[5], Restormer[6], Uformer[7] застосовують механізми самоуваги для глобального аналізу контексту. Вони досягають найвищих показників PSNR та SSIM, проте мають суттєвий недолік — надмірну параметризацію та складність інференсу.

Diffusion models: Забезпечують найвищу візуальну якість, але потребують великої кількості ітерацій генерації, що робить їх малопридатними для обробки в реальному часі.

На цьому тлі NAFNet займає проміжну позицію між класичними CNN та важкими transformer-моделями, забезпечуючи необхідну якість без значних обчислювальних витрат. Вона відмовляється від функцій активації на користь простих операцій перемноження каналів (SimpleGate[9]), що зменшує нестабільність навчання та прискорює збіжність.

Вибір NAFNet у даній роботі обґрунтований її архітектурною ефективністю, низькою параметричною складністю та високою якістю відновлення при оптимальних ресурсних витратах.

2. ДАТАСЕТ

Для навчання моделі використано модифікований набір даних, сформований на основі COCO 2017 (Common Objects in Context)[11]. COCO є одним із найпопулярніших відкритих датасетів у сфері комп'ютерного зору та містить великомасштабну колекцію реальних фотографій різної тематики.

По причині браку обчислювальних потужностей для проведення тренувань на всій тренувальній частині (120 тис. фото) цю вибірку було замінено курованим субсетом [12] (25 тис. фото).

Кожне зображення використовується двічі:

- як еталонне (чисте),
- як вхідне зашумлене, отримане шляхом накладання комбінації імпульсного та структурного шумів.

Таким чином, формуються пари (noisy image, clean image), що є стандартним підходом для задач супервізованого знешумлення.

Для генерації зашумлених зображень застосовано комбінацію декількох типів шуму:

- Гауссівський шум з нульовим середнім та дисперсією $\sigma = 0.08$;
- шум типу Salt-and-Pepper із заданою ймовірністю пошкодження пікселів;
- додаткове згладжування за допомогою гаусівського блюру.

Такий підхід дозволяє змодельовати складні реальні умови деградації якості зображень, наближені до практичних сценаріїв.

3. Архітектура нейронної мережі

3.1 Загальна концепція та принципи побудови архітектури

NAFNet (Nonlinear Activation Free Network) є спеціалізованою архітектурою для задач низькорівневої обробки зображень, зокрема знешумлення, відновлення після розмиття та суперрезолуції. Її концепція суттєво відрізняється від більшості сучасних глибоких моделей, оскільки вона відмовляється від класичних нелінійних функцій активації, замінюючи їх на більш стабільні та обчислювально ефективні механізми канального перемноження.

Основна ідея полягає в тому, що нелінійність може бути реалізована через структурну організацію операцій над каналами, а не через точкові функції активації. Це дозволяє зменшити нестабільність навчання, уникнути проблеми "згасання градієнтів" та забезпечити кращий контроль над передачею інформації між шарами.

Архітектура NAFNet побудована за принципом U-подібної структури (encoder-decoder)[10], яка поєднує багаторівневе виділення ознак зі збереженням просторової інформації за допомогою skip-зв'язків. Такий підхід є особливо ефективним у задачах відновлення, де критичне значення має збереження дрібних текстур, контурів і локальних деталей.

Важливою рисою є орієнтація архітектури на параметричну ефективність. На відміну від transformer-based моделей, що можуть містити понад 100–300 мільйонів параметрів, NAFNet демонструє конкурентну якість при значно меншій складності — у межах 2–16 мільйонів

параметрів. Це дозволяє розглядати її як практичне рішення для застосування у системах реального часу, вбудованих пристроях та мобільних платформах.

Таким чином, концептуально NAFNet є прикладом оптимізованої архітектури, яка поєднує ефективність згорткових мереж і сучасні принципи мінімізації параметричної надлишковості, забезпечуючи баланс між якістю та обчислювальною доступністю.

3.2 Структурна організація мережі та внутрішня логіка обробки

Архітектура NAFNet складається з п'яти основних функціональних частин: вхідного проєкційного шару, енкодера, центрального блоку (bottleneck), декодера та вихідного проєкційного шару. Вхідний шар здійснює первинне перетворення зображення у багатовимірний простір ознак, після чого ці ознаки поступово обробляються на різних рівнях абстракції.

Базовим елементом архітектури є блок NAFBlock, який замінює класичну комбінацію «згортка – активація – нормалізація». Він складається з Layer Normalization, pointwise convolution, depthwise convolution та механізму SimpleGate, який реалізує нелінійність через поканальне перемноження. Додатково використовується спрощений механізм каналної уваги (SCA), що підсилює значущі ознаки та зменшує вплив шумових компонентів.

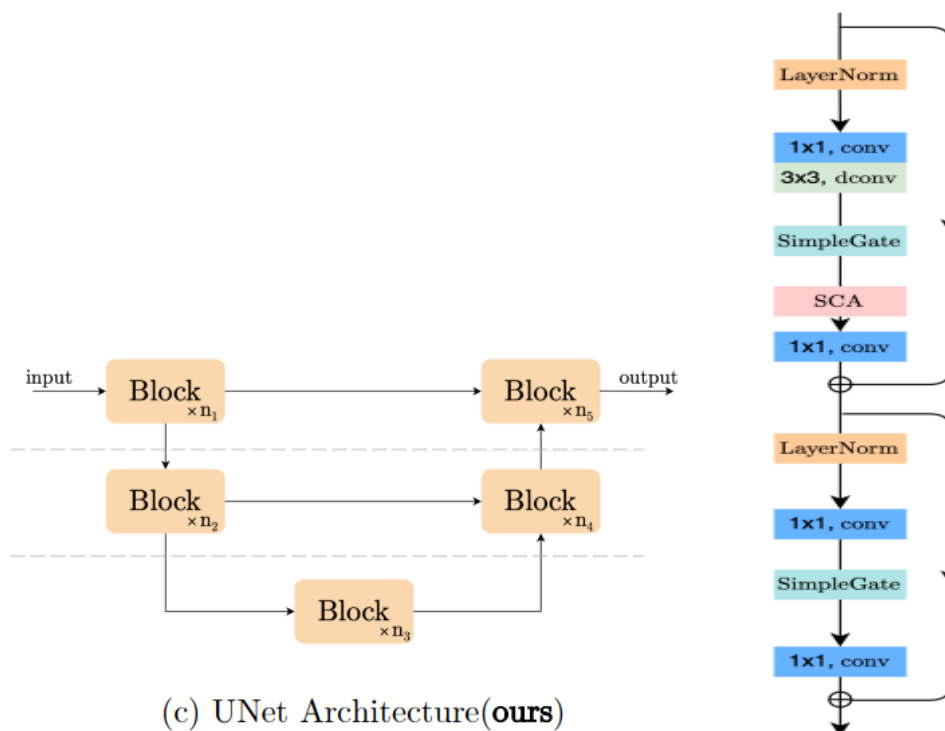


Рис. 1-2: Архітектура NAFnet (ліва), блок NAFblock (права) [9]

Енкодер (ліва частина U) відповідає за вилучення ієрархічних ознак шляхом поетапного зменшення просторової роздільної здатності та збільшення кількості каналів. На кожному рівні застосовується послідовність кількох NAFBlock, що дозволяє глибоко аналізувати локальні структури зображення. Операції downsampling не тільки зменшують розмірності, а й сприяють узагальненню інформації та формуванню високорівневих представлень.

Центральний блок (низ U) виконує роль глибокого інтегратора ознак. Саме тут відбувається консолідація інформації, зібраної з різних ділянок зображення, що дозволяє моделі формувати цілісне уявлення про структуру сцени. Завдяки цьому блок здатен коректно відновлювати глобальні закономірності, не втрачаючи локальної деталізації.

Декодер (права частина U) здійснює зворотне перетворення: поступово відновлює просторову роздільну здатність з використанням операцій апсемплінгу та skip-зв'язків. Пропускні з'єднання між відповідними рівнями енкодера і декодера дозволяють поєднати

низькорівневі просторові деталі з високорівневими семантичними ознаками, що є критично важливим для задачі знешумлення.

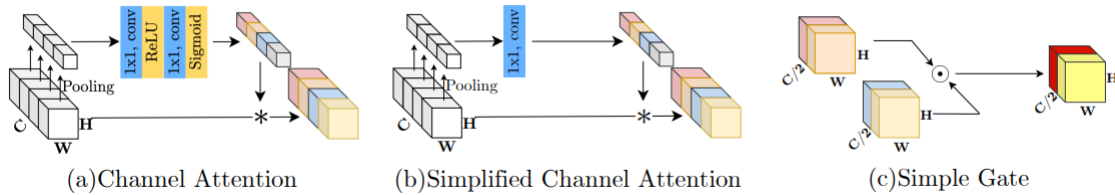


Рис. 3-5: будова модулів (a)CA, (b) SCA, (c) SimpleGate. \odot / $*$: поелементне/поканальне множення[9]

Таке поєднання компонентів забезпечує стабільну передачу градієнтів, швидке збігання та високу якість реконструкції, при цьому зберігаючи обчислювальну ефективність усієї моделі.

4. Деталі тренування

4.1 Загальна логіка навчального процесу

Процес навчання моделі знешумлення полягає у формуванні стійкої апроксимації функції, що відображає зашумлене зображення у його чисту версію. У супервізованому режимі мережа навчається мінімізувати різницю між вихідним результатом і еталонним зображенням. Для досягнення цієї мети було застосовано комплекс взаємопов'язаних компонентів, кожен з яких відіграє критично важливу роль у стабільності, швидкості та якості навчання.

У даній роботі використано наступний набір інструментів та методів:

- функція втрат L1 Loss;
- оптимізатор AdamW[13];
- стратегія зміни швидкості навчання Cosine Annealing[14] із додатковим warm-up та мультиплікативним масштабуванням;
- підпакетна обробка даних (mini-batch training) з акумуляцією градієнтів;

Кожен з цих елементів визначає конкретний аспект навчального процесу: точність відновлення, стабільність збіжності, генералізаційну здатність та обчислювальну ефективність.

4.2 Функція втрат L1 loss

L1 Loss (Mean Absolute Error) визначається як середнє абсолютне відхилення між пікселями відновленого зображення та еталонного:

$$L1 = \frac{1}{n} \sum_n |y_{pred} - y_{true}|$$

На відміну від L2 Loss, який квадратично штрафує великі похибки, L1 Loss є більш стійкою до викидів та локальних аномалій, що є важливою властивістю у задачі знешумлення. Вона сприяє збереженню дрібних деталей та запобігає надмірному згладжуванню, яке часто виникає при використанні квадратичної похибки.

Практичними перевагами L1 Loss є:

1. стабільна поведінка на шумних вибірках;
2. краща реконструкція текстур;
3. менша чутливість до поодиноких спотворень;
4. збереження чітких контурів об'єктів.

4.3 Оптимізатор AdamW

AdamW[13] є модифікованою версією оптимізатора Adam, у якій регуляризаційний компонент вагового згасання (weight decay) відокремлений від механізму адаптивної корекції градієнтів. Це дозволяє уникнути небажаного впливу регуляризації на швидкість навчання.

Основні характеристики AdamW:

1. адаптивне налаштування learning rate для кожного параметра;
2. експоненціальні ковзні середні першого та другого моментів градієнта;
3. стабільність при великих навчальних вибірках;
4. покращена узагальнювальна здатність у порівнянні зі стандартним Adam.

У задачах відновлення зображень AdamW забезпечує швидке збігання без втрати якості, що робить його оптимальним вибором для навчання архітектури NAFNet.

4.4 Планувальник швидкості навчання

Оригінальна стратегія Cosine Annealing[14] передбачає плавне зменшення learning rate відповідно до косинусоїдальної функції, із мультиплікативним збільшенням тривалості циклу спадання. Для покращення адаптивності цієї стратегії додатково було додане мультиплікативне масштабування – тобто з кожним циклом не тільки збільшуватиметься тривалість спадання, а й зменшується «розмір» косинусоїдальної функції, з якої це спадання відбувається. Додатково для конфігурації доступний розігрів (warm-up), котрий дозволяє зберегти стабільність на початковій фазі.

Переваги такої стратегії:

1. стабілізація початкової фази навчання;
2. Стійкість до локальних мінімумів – суттєві коливання швидкості навчання успішно виводитимуть мережу з них;
3. Можливість балансувати між збіжністю тренування та здатністю мережі «аналізувати нові ділянки функції втрат» - за рахунок мультиплікативного масштабування, множник котрого можна змінювати у тому числі посеред тренування – цим самим досягаючи стабільніших мінімумів та, відповідно, кращої генералізації.

4.5 Обробка підпакетами та акумуляція градієнтів

Навчання здійснюється у режимі mini-batch, коли зображення обробляються невеликими пакетами. Це дозволяє імітувати великий розмір пакета – тобто не жертвувати стабільністю оцінки градієнтів - переносючи балансування на площину ресурсів та часу.

Через обмеження об'єму відеопам'яті застосовується акумуляція градієнтів, при якій градієнти обчислюються для кількох підпакетів поспіль і лише після цього виконується оновлення ваг «для всього пакета». Це дозволяє емулювати ефект великого batch size без перевищення доступних ресурсів.

4.6 Параметри та інші деталі

Кількість застосованих епох – 10. Розмір емульованого пакета – 128. Розмір картинок після пайплайну конвертацій – 256×256 . Швидкість навчання = 0.001, фактор затухання $\text{var} = 0.001$. Dropout застосований не був.

5. Отримані результати

Для оцінки загальної продуктивності моделі було використано метрику порівняння зображень Multi-Scale Structural Similarity Index (ms-ssim), котра при вимірюванні враховує освітлення, контраст та власне структуру зображення. У деякому плані, значення цієї метрики показують схожість картинок з людської, персептивної точки зору.

Для зашумлення використовувався ось такий пайплайн Pytorch-трансформацій:

```
noise_transform = v2.Compose([
    v2.ToImage(),
    v2.Resize(size=(img_size, img_size)),
    v2.ToDtype(torch.float32, scale=True),
    v2.GaussianNoise(mean=0, sigma=0.08),
    BatchSaltAndPepper(salt_prob=0.05, pepper_prob=0.05),
    v2.GaussianBlur(kernel_size=5, sigma=(0.5, 1.5))
])
```

де `BatchSaltAndPepper` – вручну імплементована трансформація з інтуїтивно зрозумілими параметрами.

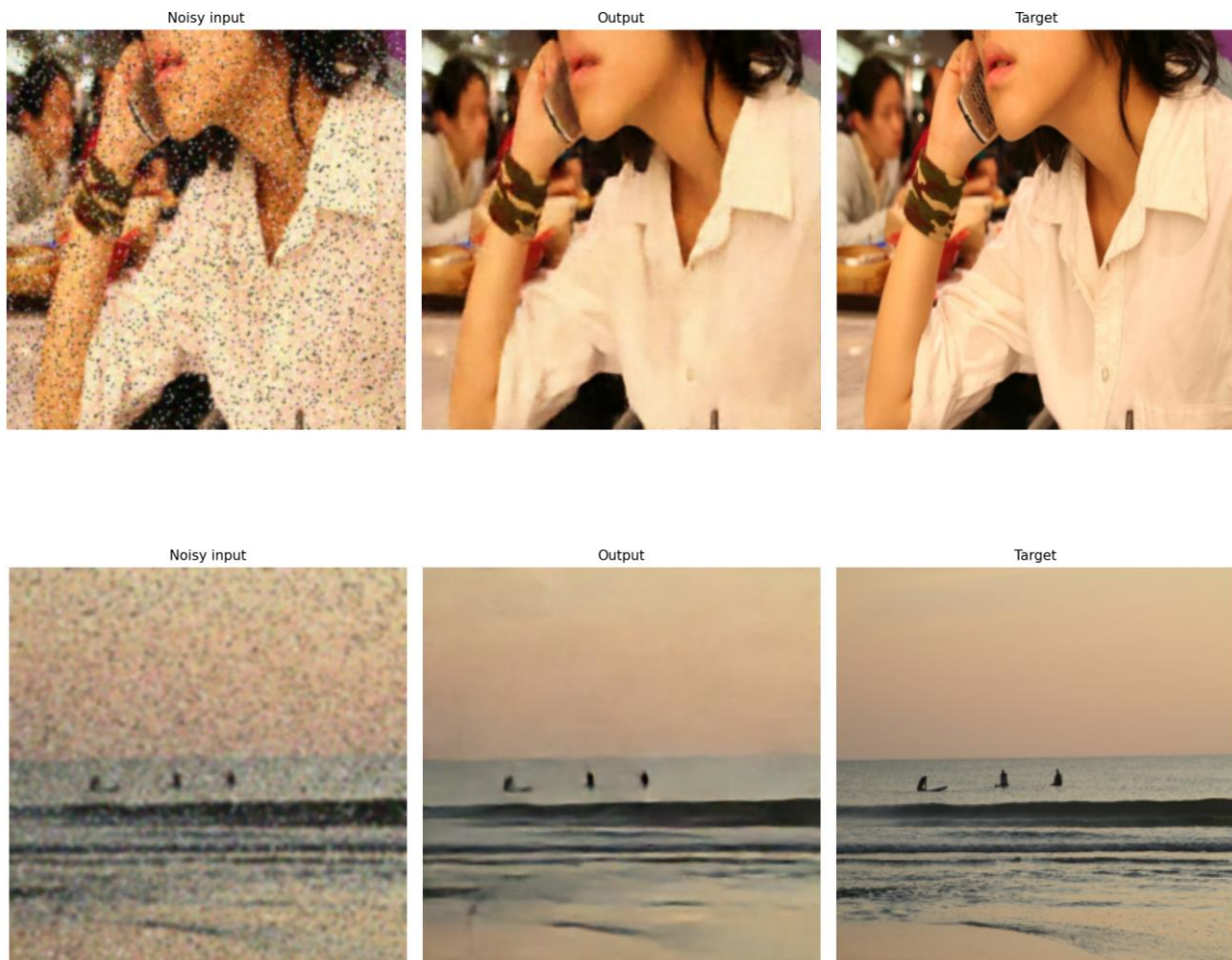
В результаті цього пайплайну отримувалися картинки із середнім (по всій тренувальній вибірці) `ms-ssim` (порівняння з оригіналом) рівним 78.5%, тобто втрати інформації при зашумленні перевищують 20%.

Натренована модель навчилася компенсувати шум, зменшуючи втрату інформації із 20+% до всього 5%, тобто останні моделі видавали в середньому `ms-ssim` = 95%.

```
Loading pretrained model, optimizer & scheduler state dicts...
all the dicts set up successfully.
testing done. input ms-ssim: 78.256%; output ms-ssim: 95.072%
```

Рис. 2: Результати роботи натренованої моделі на тестовій вибірці MS-COCO

Нижче наведено ілюстрації роботи результуючого NAFnet на тестовому сеті:



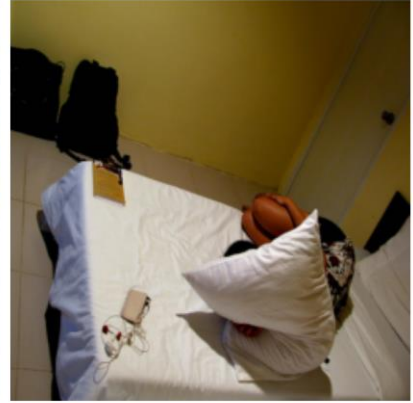
Noisy input



Output



Target



Noisy input



Output



Target



Noisy input



Output



Target





Рис. 7-24: Ілюстрації роботи NAF-net на зображеннях тестувального сету ms-COCO. Зліва направо: вхідне зображення, знешумлене зображення, оригінал

Досягнуто значної якості відновлення. Проте все ще спостерігається втрата інформації – зображення здаються замиленими. Це спровоковано у тому числі тим, що імпульсні шуми із послідовним зашумленням структури призводять до сильного викривлення цієї структури, провокуючи цим неможливість її відновлення у чітких деталях.

ВИСНОВКИ

У даній науково-практичній роботі було досліджено та реалізовано підхід до знешумлення цифрових зображень із використанням глибокої нейронної мережі архітектури NAFNet, орієнтованої на задачі низькорівневої обробки. Основною метою було досягнення оптимального балансу між якістю відновлення зображень та обчислювальною ефективністю моделі.

У ході виконання роботи було проаналізовано сучасні підходи до задачі знешумлення, включаючи класичні фільтраційні методи, CNN-архітектури, transformer-моделі та дифузійні підходи. Проведений аналіз показав, що надпараметризовані моделі, хоча й демонструють високу якість, є малопридатними для практичного застосування в умовах обмежених ресурсів, що обґрунтовує вибір архітектури NAFNet як ефективної альтернативи.

Для навчання моделі було використано модифікований субсет COCO 2017 з синтетично змодельованими шумами складної природи, що наближує експеримент до реальних сценаріїв деградації зображень. Навчання проводилося із застосуванням L1 Loss, оптимізатора AdamW акумуляції градієнтів з емуляцією пакетів та адаптивної стратегії зміни швидкості навчання.

Експериментальні результати показали, що модель змогла суттєво підвищити структурну подібність між відновленими та оригінальними зображеннями: середній показник ms-ssim зріс із 78,5% для зашумлених зображень до 95% після обробки нейронною мережею. Це свідчить про високу ефективність обраного підходу, хоча в окремих випадках спостерігалось незначне розмиття та втрата дрібних деталей, зумовлені складним характером комбінованих шумів.

Таким чином, результати роботи підтверджують, що архітектура NAFNet є перспективною для практичного застосування в задачах знешумлення зображень, особливо в умовах обмежених апаратних ресурсів.

ЖИТЕПАТҮПА

1. Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7), 3142–3155. <https://doi.org/10.1109/TIP.2017.2662206>
2. Mao, X.-J., Shen, C., & Yang, Y.-B. (2016). Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems (NeurIPS 2016)* (pp. 2802–2810). <https://papers.nips.cc/paper/6172-image-restoration-using-very-deep-convolutional-encoder-decoder-networks-with-symmetric-skip-connections.pdf>
3. Zhang, K., Zuo, W., & Zhang, L. (2018). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9), 4608–4622. <https://doi.org/10.1109/TIP.2018.2839891>
4. Zhang, K., Zuo, W., Gu, S., & Zhang, L. (2017). Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3929–3938). <https://doi.org/10.1109/CVPR.2017.300>
5. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: Image restoration using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCV W)*, 1833–1844. <https://doi.org/10.1109/ICCVW54120.2021.00210>
6. Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5728–5739). <https://doi.org/10.1109/CVPR52688.2022.00564>
7. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., & Li, H. (2022). Uformer: A general U-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 17662–17672). <https://doi.org/10.1109/CVPR52688.2022.01716>
8. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv*. <https://doi.org/10.48550/arXiv.2006.11239>
9. Chen, L., Chu, X., Zhang, X., & Sun, J. (2022). Simple baselines for image restoration (NAFNet). In *European Conference on Computer Vision (ECCV 2022)*. https://doi.org/10.1007/978-3-031-19803-8_5
10. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
11. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV 2014)* (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48
12. Samet, N., Hicsonmez, S., & Akbas, E. (2020). HoughNet: Integrating near and long-range evidence for bottom-up object detection. In *European Conference on Computer Vision (ECCV 2020)*. https://doi.org/10.1007/978-3-030-58542-6_34; <https://www.kaggle.com/datasets/trungit/coco25k/data>
13. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR 2019)*. <https://doi.org/10.48550/arXiv.1711.05101>
14. Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR 2017)*. <https://doi.org/10.48550/arXiv.1608.03983>