Одеський національний університет імені І. І. Мечникова

Факультет математики, фізики та інформаційних технологій

Кафедра оптимального керування і економічної кібернетики

Дипломна робота

бакалавра

на тему: «Аналіз ефективності методів віднімання фону»

«Natural Image Matting»

Виконав: студент денної форми навчання
спеціальності 113 Прикладна математика

Підгорний Богдан Валерійович

Керівник канд. техн. наук, доц. Мороз В.В.
(науковий ступінь, вчене звання, прізвище та ініціали, підпис)
Рецензент канд. техн. наук, доц. Мазурок І.Є.
(науковий ступінь, вчене звання, прізвище та ініціали)

Рекомендовано до захисту:

Протокол засідання кафедри

№ ___ від _____ р.

Завідувач кафедри

_____ _____
(підпис) (прізвище, ініціали)

Захищено на засіданні ЕК №_____

протокол № ____ від _____ р.

Оцінка_____/___ ___/_____
(за національною шкалою, шкалою ECTS, бали)

Голова ЕК

_____ _____
(підпис) (прізвище, ініціали)

Одеса – 2022

Зміст

Вступ

Задача віднімання фону - це важлива і актуальна проблема сьогодення, методи вирішення якої активно розвиваються, науковці працюють з огляду на велику конкуренцію, що прискорює винаходження нових підходів. Віднімання фону є дуже зручним та сьогодні майже незамінним інструментом у багатьох сферах, а саме: редагування фото та відео, створення фільмів, та їх пост-продакшн, комп'ютерний зір та таке інше. Розвиток комп'ютерного зору є важливим в першу чергу різноманіттям областей де він використовується і так, як віднімання фону є дуже важливим аспектом до нього, то будуючи нові алгоритми і, рухаючи цей напрямок, оптимізуються процеси у багатьох напрямках. Також останнім часом актуальність проблеми для людей різко зросла, тому що світ у період пандемії COVID-19 почув необхідність у програмах, що використовують реалізації алгоритмів віднімання фону, а саме – такі програми як Skype, Zoom, та інші. Для створення атмосфери офісу, або для конфіденційності люди почали активно використовувати заміну фону під час мітингів та приватних дзвінків. Учасники онлайн-заходів з різними пристроями почали масово використовувати програми по заміні фону і це показало головні проблеми реалізацій сьогодення, а саме: швидкість роботи та використовування великої кількості оперативної пам'яті.

Хоча за останні кілька років підходи, засновані на глибокому навчанні, досягли видатних результатів у вирішенні проблеми віднімання фону зображення. Однак все ще є два недоліки, які перешкоджають їх широкому застосуванню: залежність від наданих користувачем даних про контур переднього та заднього плану зображення (trimap), і великі розміри моделей. У цій роботі пропонується алгоритм віднімання фону без трімапів (trimap-free) та з модифікованою моделлю нейронної мережі. За допомогою полегшенного базового блоку згортки будується двоетапна структура:

1. мережа сегментації (SN) призначена для захоплення достатньої кількості семантики та класифікація пікселів на невідомі, передній фон та задній фон;
2. matting Refine Network (MRN) спрямований на отримання детальної інформації про текстуру та регресію точних значень альфа. З пропонованим міжрівневим модулем злиття (CFM), SN може ефективно використовувати багатомасштабні особливості(feature) з меншими витратами на обчислення. Ефективний нелокальний модуль уваги (ENA) в MRN може ефективно моделювати релевантність між різніми пікселями та допомагає регресувати високоякісні значення альфа. Використовуючи ці техніки, створено надзвичайно легку модель, яка досягає порівнянної продуктивності з параметрами 1% (344k) великих моделей на популярних тестових бенчмарках віднімання фону зображень.

Але без наданої користувачем trimap є дві основні задачі для автоматичного віднімання фону:

1. визначити передній план, задній план і невідомі області на зображенні;
2. передбачити точні значення альфа невідомих регіонів.

Перша — це класична класифікаційна задача, друга задача — типове завдання регресії. Розв'язання єдиною структурою, таким чином, пропонується двоетапна end-to-end deep-learning структура для легкого процесу віднімання фону без trimap. Перший етап розроблений щоб охопити достатню семантику для визначення переднього плану, фону та невідомих регіонів. Тоді на основі першого етапу другий етап може дати більш точні значення альфа-каналу невідомих регіонів.

- ***Об'єктом дослідження*** є процес обробки зображень.
- ***Предметом*** є алгоритм для вирішення задачі віднімання фону.
- ***Мета дослідження*** - розробка алгоритму віднімання фону без трімапів (trimap-free) та з модифікованою моделлю нейронної мережі.

Розділ 1. Огляд методів та алгоритмів віднімання фону

Багато сучасних алгоритмів мають на меті розв'язати рівняння матування, розглядаючи його як проблему кольору після підходів вибірки або поширення. Віднімання зображень на основі зразка передбачає, що справжні кольори переднього плану та фону невідомого пікселя можуть бути отримані з відомих зразків переднього плану та фону, які знаходяться поблизу цього пікселя. Методи, які дотримуються цього припущення, включають описані нижче.

1.1 Bayesian matting

Для подальшого розвитку будемо вважати, що вхідне зображення вже розділено на три області: «фон», «передній план» і «невідомий», при цьому фон і області переднього плану були окреслені консервативно. Таким чином, мета алгоритму полягає в розв'язанні кольора переднього плану F, кольора фону B і непрозорість α враховуючи спостережуваний колір C для кожного пікселя в межах невідомої області зображення. Оскільки F, B і C мають по три колірних канали, у нас є проблема з трьома рівняннями і сімома невідомі. Подібно до методів Рузона та Томасі, проблема вирішується частково шляхом побудови переднього і заднього плану ймовірності розподілу з даної околиці. Метод, однак, використовує безперервно ковзне вікно для сусідства визначення, рухається всередину від переднього плану та фону та використовує найближчі обчислені F, B та α значення (на додаток до цих значень із «відомих» регіонів) в побудові орієнтованих гауссових розподілів. Далі підхід формулює проблему обчислення матових параметрів у чітко визначеному байєсіанському каркасі і вирішує його за допомогою максимальної апостеріорної (MAP) техніки.

1.2 Shared sampling matting

Техніка альфа-матування в реальному часі заснована на Фундаментальному спостереженні, що пікселі в невеликій околиці часто мають дуже подібні значення для своїх істинних (α,F,B). З цього випливає, що: початкова колекція вибірок, зібраних сусідніми пікселями, відрізняється лише невеликою кількістю елементів; та зазвичай вибираються близькі пікселі однакові або дуже схожі пари для їх найкращого переднього плану та кольорів фону. Це призводить до висновку, що виконання альфа-матового обчислення для кожного пікселя незалежно від його сусідів призводить до великої кількості зайвих робіт, яких можна безпечно уникнути без шкоди якості віднімання фону. Насправді, можна досягти прискорень до двох порядків, але при цьому отримати високоякісні результати. Ця техніка приймає в якості вхідних даних зображення (або відеопослідовність) і відповідну(-і) відображення(-и) і складається з наступних кроків:

1. Розширення відомих регіонів: екстраполює «відомий передній план» та «відомий фон» у «невідомі» регіони;

2. Відбір зразка та обчислення фону: намагається визначити для кожного пікселя в невідомих областях найкращу пару зразків переднього плану та фону. Він також обчислює альфа-мат з отриманих зразків;

3. Локальне розгладжування: локально розгладжує отриманий мат, зберігаючи свої відмінні риси.

1.3 Spase coding

Він вирішує дві основні проблеми, з якими стикалися попередні алгоритми на основі вибірки. По-перше, існуючі підходи, засновані на вибірці, зазвичай покладаються на певні просторові припущення при зборі зразків з відомих регіонів і, отже, їх ефективність погіршується, якщо основні припущення незадовільні. Друга

проблема полягає в тому, що якість результату віднімання визначається добротністю однієї пари зразків, що викликає помилки коли методи на основі вибірки не можуть вибрати найкращі пари.

Віднімання поширенням зображення працює шляхом поширення відомого альфа-значення між відомими локальними зразками переднього плану та фону до невідомих пікселів. Приклади наведені нижче.

1.4 Poisson matting

На відміну від попередніх методів, які оптимізують піксель кольори альфа, фону та переднього плану статистично, метод діє безпосередньо на градієнті зображення. Це зменшує помилку, викликану неправильною класифікацією колірних зразків у зображенні. Матування Пуассона оцінює градієнт матування від зображення, потім відновлює мат, розв’язуючи рівняння Пуассона. Формулювання базується на припущенні, що інтенсивність змінюється на передньому та задньому плані. Пропонуються глобальні матування Пуассона, напівавтоматичний підхід до приблизного матування з градієнта зображення, заданого користувачем. Передній план і кольори фону можна буде більш надійно оцінити. Що ще важливіше, коли глобальне матування Пуассона не виходить високоякісні мати завдяки складному фону, вводиться локальне матування Пуассона, яке маніпулює безперервним градієнтним полем в місцевому регіоні. Місцеве матування Пуассона забезпечує взаємодію людини у петлю «витягування матового». У більшості випадків градієнти зображення викликані кольорами переднього плану та фону, візуально розрізняються локально. Знання від користувача можна ефективно інтегрувати в матування Пуассона за допомогою набору інструментів, які діють на градієнтне поле матового. В результаті метод може підтримуватися безперервності тонких довгих ниткоподібних форм об'єктів переднього плану у складній сцені. Крім того, після локальних операцій користувачі можуть легко вносити зміни в градієнтне поле.

1.5 Geodesic matting

1. Метод заснований на зважених функціях відстані (геодезичних), завдяки чому вирішується геометричне рівняння Гамільтона-Якобі першого порядку за оптимальний для обчислень лінійний час. Це робить запропонований фреймворк віднімання фону для інтерактивної обробки зображень і відео.

2. Метод дає дуже хороші, найсучасніші результати, з дуже небагато і інтуїтивно зрозумілих користувачів надають каракулі і дуже прості атрибути, що визначають ваги при обчисленні відстані. Часто використовуємо лише пару грубих каракулів нерухомого зображення (один для переднього плану і один для фону).

3. Метод використовує як розташування каракулів, так і статистику пікселів, позначених користувачем, використовуючи, таким чином, усю надану ним/нею інформацію.

4. Це стосується великого класу природних даних, а тому й дозволяє уникнути офлайн-навчання, воно не обмежується попереднім спостереженням та засекреченими класами, а також доступністю достовірних даних і сегментованих вручну даних.

5. Він також може обробляти динамічний фон у відео перетин цікавих об'єктів.

6. Структура є загальною, тому додаткові атрибути можуть включатися до ваг для геодезичних відстаней, якщо це потрібно для конкретного типу даних.

1.6 Spectral matting

До методів спектральної сегментації, вдаються обчислення дійсних власних векторів як наближення, необхідного для перетворення NP-повної проблеми на вирішувану. Навпаки, шукається не розрізнений розділ зображення, а скоріше спроба відновити дробове покриття переднього плану на кожному пікселі. Зокрема, дійсно цінні компоненти матування отримуються через лінійне перетворення найменших власних векторів матриці Лапласа, введеної Левіним.

Після отримання ці матуючі компоненти служать будівельними блоками для побудови повних матів переднього плану.

1.7 Close-form matting

Вводиться функція вартості на основі локальних припущень про гладкість для кольорів F і B переднього плану та фону та можна показати, що з отриманого виразу можна аналітично виключити F і B, отримуючи квадратичну функцію вартості в α. Альфа-мат, отриманий за методом, є глобальним оптимумом цієї функції вартості, яку можна отримати, розв’язавши розріджену лінійну систему. Оскільки підхід обчислює α безпосередньо і не вимагає надійних оцінок для F і B, а дивно мала кількість введених даних користувача (наприклад, розріджений набір каракулів) часто достатньо для отримання високої якості мату. Крім того, закрита формула дозволяє зрозуміти та передбачити властивості шляхом ретельного вивчення власних векторів розрідженої матриці пов’язаних з матрицею, що використовуються в алгоритмах сегментації спектрального зображення.

1.8 Fuzzy connectedness matting

FuzzyMatte значно скорочує час між кожною ітерацією інтерактивного циклу, ефективно інтегруючи новий вхід з оцінкою мату з попередньої ітерації. FuzzyMatte нащадником нечіткої сегментації, які використовуються для визначення часткового об’єму від кілька слоїв на одному зображенні. Нехай дано зображення і початковий введений користувачем (каракулі або тримап): спочатку обчислюємо для кожного пікселя p з невідомим альфа-матом його нечітку зв'язаність (FC) з відомим переднім і заднім планом. FC – це концепція, яка ефективно фіксує нечітку «зв’язаність» (суміжність і подібність) між елементами зображення. FC для кожного пікселя можна обчислити шляхом пошуку

найсильнішого зв’язаного шляху до відомого переднього і заднього плану. Значення альфа пікселя потім можна розрахувати з FC. Коли стають доступними додаткові дані, FuzzyMatte не вимагає повторного обчислення FC всіх пікселів шляхом пошуку найсильнішого зв’язаного шляху. Замість цього він розбиває пікселі на три підмножини на основі їхніх старих значень FC і лише невелика кількість пікселів в одному з трьох підмножин вимагають пошуку найсильнішого шляху зв'язку. Всі інші пікселі просто повторно використовують попередньо оцінений FC.

Однак надмірна залежність від інформації про колір може призвести до артефактів у зображеннях, де розподіл кольорів переднього та фонового плану перекриваються. Таким чином, останніми роками було розроблено кілька підходів на основі глибокого навчання, головним з них є двоступенева мережа, що складається з етапу кодера-декодера та етапу уточнення.

1.9 Двоступенева мережа, що складається з етапу кодера-декодера та етапу уточнення, розробленого Xu et al

Метод використовує глибоке навчання для безпосереднього обчислення альфа-мату за вхідним зображенням і treemap. Замість того, щоб покладатися в першу чергу на інформацію про колір, мережа може навчитися знаходити структуру в альфа-матах. Наприклад, волосся і хутро (які зазвичай вимагають матування) володіють сильними структурними і текстурними візерунками. Інші випадки, які потребують матування (наприклад, краї предметів, області оптичного або розмиття в русі, або напівпрозорі області) майже завжди мають спільну структуру або альфа-профіль, який можна очікувати. Поки низькорівневих функцій не буде захоплюючи цю структуру, глибокі мережі ідеально підходять для її представлення. Дана двоступенева мережа включає в себе етап кодер-декодер, за яким слідує невелика залишкова мережа для уточнення і включає втрату нового складу на додаток до втрати на альфі.

Розділ 2. Математичні моделі віднімання фону

2.1 Модель на основі глибокого навчання

Однією з перспективних моделей є модель на основі глибокого навчання.

Перший етап мережі - глибокий кодер-декодер мережі, який досяг успіхів у багатьох інші задачах комп’ютерного зору, таких як сегментація зображення, прогнозування границь та заповнення отворів.

Структура мережі будується на основі вхіднмх даних: зображення та відповідна тримап, які об’єднані вздовж розміру каналу, що призводить до 4-канального введення. Вся мережа складається з мережі кодерів і декодерыв. Вхід до мережі кодера є, перетворені в асоціації об’єктів зі зниженою дискретизацією, згорткові шари та шари максимального об’єднання. Мережа декодера, у свою чергу, використовує наступні шари видалення пулу, які змінюють максимальну операцію об'єднання та згорткові рівні щоб збільшити вибірку карт функцій і отримати бажаний результат, альфа-мата. Зокрема, мережа кодерів має 14 згорткових шарів і 5 шарів максимального об’єднання. Для мережі декодера використовується менша структура, ніж для мережі кодера для зменшення кількості параметрів і прискорення тренувального процесу. Зокрема, мережа декодеру має 6 згорткових шарів, 5 шарів роз’єднання, за якими слідує останній рівень альфа-прогнозування.

Втрати: мережа використовує дві втрати. Перша втрата називається втратою альфа-прогнозування, яка є абсолютною різницею між значеннями альфа істини та передбачуваними значеннями альфа для кожного пікселя. Однак через недиференційовану властивість абсолютних значень використовується наступна функція втрат, щоб наблизити її.

$$\mathcal{L}_{\alpha}^{i} = \sqrt{(\alpha_p^i - \alpha_g^i)^2 + \epsilon^2}, \quad \alpha_p^i, \alpha_g^i \in [0, 1].$$

Де α_p^i є вихідним результатом шару передбачення в пікселі i порогове значення від 0 до 1. α_g^i – це нижнє порогове значення альфа в пікселі i. ϵ – дуже маленьке число близьке до 10^{-6} в експеріментах.

$$\frac{\partial \mathcal{L}_{\alpha}^{i}}{\partial \alpha_p^i} = \frac{\alpha_p^i - \alpha_g^i}{\sqrt{(\alpha_p^i - \alpha_g^i)^2 + \epsilon^2}}.$$

Друга втрата називається композиційною втратою, тобто абсолютна різниця між найменшими істиними кольорами RGB і передбачуваними кольорами RGB, складеними з істиним найменшим значенням переднього плану, найменшим істиним значенням фону і передбаченим альфа. Аналогічно маємо його наближення за допомогою наступної функції втрат.

$$\mathcal{L}_{c}^{i} = \sqrt{(c_p^i - c_g^i)^2 + \epsilon^2}.$$

Де c позначає канал RGB, p позначає зображення складається з передбаченої альфа-версії, а g позначає зображення складається з нижнього істинного значення альфа. Композиційна втрата є композиційною операцією, що призводить до більш точних альфа-прогнозів.

Загальна втрата є зваженою сумою двох окремих втрат, тобто

$$\mathcal{L}_{overall} = w_l \cdot \mathcal{L}_{\alpha} + (1 - w_l) \cdot \mathcal{L}_{c},$$

Де w_l в експерименті встановлено значення 0,5. Встановлюються додаткові ваги для двох типи втрат відповідно до розташування пікселів, які можуть допомогти мережі

приділяти більше уваги важливим сферам. Зокрема, при $w_i = 1$ якщо піксель i входить до невідомого регіону трімапи, а $w_i = 0$ навпаки.

Реалізація: хоча навчальний набір даних має 49 300 зображень, маємо тільки 493 унікальних об'єкта. Щоб уникати перенавчання, а також для ефективнішого використання навчальних даних, використовуються кілька стратегій навчання. Спочатку випадковим чином обрізаються пари 320×320 (зображення, обробка) з центром пікселів у невідомих регіонах. Це збільшує простір для вибірки. Далі також обрізаються навчальні пари різних розмірів (наприклад, 480×480, 640×640) та змінюємо їх розмір до 320×320. Це робить метод більш надійним для масштабування і допомагає мережі краще вивчати контекст і семантику. Перевертання виконується випадковим чином на кожній навчальній парі. Трімапи випадковим чином розширюються від істиного найменшого значення альфа-мату, що допомагає моделі бути більш надійною при розміщенні тримапу. Нарешті, навчальні вхідні дані відтворюються випадковим чином після кожної епохи навчання. Частина кодера мережі ініціалізується за допомогою перших 14 згорткових шарів ВГГ-16 (14-й шар є повністю шаром "fc6", який можна трансформувати до згорткового шару). Так як мережа має 4 канали введення, ініціалізується один додатковий канал першого рівня. Всі параметри декодера ініціалізуються випадковими величинами Xavier. Під час тестування, зображення та відповідна обробка з'єднанаются у єдиний набір вхідних даних. Прямий прохід мережі виконується для виведення передбачення альфа-мату.

2.2 Модель Sampling/Segmentation Algorithm

Тримап T сегментує вхідне зображення (або відеокадр) у три піксельні області, що не перекриваються: відомий передній план T_f, відомий фон T_bта невідома частина T_u. Ідея розширення відомих областей означає використання спорідненості сусідніх пікселів для зменшення розміру невідомих областей. Таким чином, нехай $D_{image}(p, q)$, $D_{color}(p, q)$ будуть, відповідно, відстань між двома пікселями p і q. Процес розширення

складається з перевірки для кожного пікселя $p \in T_u$ якщо існує піксель $q \in T_r (r = \{f, b\})$ такий, що $D_{image}(p, q) \leq k_i$, $D_{color}(p, q) \leq k_c$, та $D_{image}(p, q)$ мінімальне для p. У такому випадку піксель p помічається як той що відноситься до регіону T_r через його спорідненність з пікселем $q \in T_r$. Значення параметра k_i залежить від розміру невідомого регіона: так більші зображення можуть вимагати більших значень параметра k_i. Помічено, що $k_i = 10$ пікселів та k_c = 5/256 пунктів (вимірюється як евклідова відстань в одиничному кубі RGB) дають хороші результати для типових зображень.

Мінімізування обсягу роботи, пов'язаної з пошуком найкращих пар зразків переднього плану та фону для набору сусідніх пікселів, відбувається за допомогою використання їх високої спорідненості.

2.2.1 Збір зразків

На цьому етапі кожний піксель $p \in T_u$ шукає можливі зразки у передньому плані та задньому плані з k_g лінійних сегментів починаючи з p. Ці сегменти ділять зображення на k_g розрізнених секторів, що містять рівні плоскі кути. Нахил першого відрізка лінії, пов'язаного з p, визначається як початкова орієнтація $\theta \in [0; \frac{\pi}{2}]$ та вимірюється по відношенню до горизонтальної лінії. Такий кут приймає різне значення для кожного пікселя $q \in T_u$ у вікні 3×3. Орієнтація на інші відрізки задані кутовим приростом $\theta_{inc} = \frac{2\pi}{k_g}$. Таким чином, p має знайти свою найкращу пару серед, щонайбільше, k_g^2 пар зразків.

Маємо нову цільову функцію, яка поєднує фотометричні, просторові та ймовірнісні елементи, щоб вибрати найкращий зразок пари для пікселя з початкового набору k_g^2 пар. Запропонований підхід є першим комплексно поєднавшим всі ці аспекти. Таким чином, нехай f_i , b_j будуть парою переднього плану і зразки фону, кольори яких є F^i , B^j. Далі виводиться цільова функція для визначення найкращої пари вибірки для кожного пікселя $p \in T_u$. Перш ніж описувати цю функцію, укажемо її необхідні будівельні блоки (мінімізація хроматичного спотворення та статистика простору зображення).

2.2.2 Мінімізація хроматичного спотворення

Мінімізація моделюється хроматикою спотворення M_p, значення якого має бути невеликим для хорошої пари кандидатів кольорів:

$$M_p(F^i, B^j) = \|C_p - (\hat{\alpha}_p F^i + (1 - \hat{\alpha}_p) B^j)\|$$

де C_p- колір p, та $\widehat{\alpha_p}$ - оцінене значення альфа для p, отримане за допомогою проекції колірного простору C_p на пряму, визначену як F^i , B^j.

Хороша пара зразків повинна мінімізувати залишок найменших квадратів, визначений терміном спорідненості сусідства:

$$N_p(\mathbf{f}_i, \mathbf{b}_j) = \sum_{q \in \Omega_p} M_q(F^i, B^j)^2$$

де Ω_p— околиця пікселя p, що складається з усіх пікселів у вікні 3 × 3 з центром у точці p, а M_q — це оператор, вичислений у пікселі q.

2.2.3 Статистика простору зображення

Нехай $D_p(s) = D_{image}(s, p) = ||s - p||$ буде відстанню простору зображення від зразка s до поточного пікселя p. Визначаємо енергію $E_p(s)$ щоб досягти зразка переднього або фонового плану поточного пікселя p як квадратований інтеграл шляху ∇I вздовж відрізку L простору зображення, що з'єднує p і s:

$$E_p(\mathbf{s}) = \int_L \|\nabla I \cdot d\mathbf{r}\|^2 = \int_p^{\mathbf{s}} \left\| \nabla I \cdot \left(\frac{\mathbf{s} - p}{\|\mathbf{s} - p\|} \right) \right\|^2 .$$

Енергія $E_p(s)$ прямо пропорційна добутку прогнозованої довжині ∇I на нормований напрямок інтегрування $(s - p)$. Таким чином, якщо лінійний шлях від s до p перетинає регіони зображення, де $|\nabla I|$ велике (наприклад, по краях зображення), більша енергія

буде потрібна для досягнення s. Оцінка ймовірності належності p до переднього плану відповідно до енергії $E_p(s)$ можна отримати як:

$$PF_p = \frac{min_j(E_p(\mathbf{b}_j))}{min_i(E_p(\mathbf{f}_i)) + min_j(E_p(\mathbf{b}_j))}.$$

Таким чином, якщо мінімальна енергія, необхідна для досягнення переднього плану зразку набагато нижча, ніж мінімальна необхідна енергія для отримання фонового зразку, то PF_p буде близьким до одиниці, тобто піксель p має високу ймовірність приналежності до переднього плану. Інтуїтивно значення альфа $\widehat{\alpha_p}$ (обчислюється з f_i, b_j) можемо співвідносити з ймовірністю PF_p пікселя p, що належить передньому плану. Таким чином, хороша пара зразків також повинна мінімізувати функцію A_p

$$A_p(\mathbf{f}_i, \mathbf{b}_j) = PF_p + (1 - 2\,PF_p)\,\hat{\alpha}_p.$$

Дійсно, для даної пари зразків з (f_i, b_j), коли $PF_p = 0$, $A_p(f_i, b_j) = \widehat{\alpha_p}$. Таким чином, мінімізація $A_p(f_i, b_j)$ також мінімізує значення $\widehat{\alpha_p}$. Так само, коли $PF_p = 1$, $A_p(f_i, b_j) = (1 - \widehat{\alpha_p})$, мінімізація $A_p(f_i, b_j)$ максимізує $\widehat{\alpha_p}$. В кінці кінців при $PF_p = 0.5$, $A_p(f_i, b_j) = 0.5$ та значення $\widehat{\alpha_p}$ не має ефекту на мінімізацію. $A_p(f_i, b_j)$ буде використовуватися як один з термінів в кінцевій цільовій функції.

Цільова функція:

Отримана цільова функція що поєднує фотометричну та просторову спорідненість, а також імовірнісна інформація для вибору хороших пар вибірок фону та переднього плану може бути виражена як:

$$g_p(\mathbf{f}_i, \mathbf{b}_j) = N_p(\mathbf{f}_i, \mathbf{b}_j)^{e_N} A_p(\mathbf{f}_i, \mathbf{b}_j)^{e_A} D_p(\mathbf{f}_i)^{e_f} D_p(\mathbf{b}_i)^{e_b}.$$

$N_p(f_i, b_j)$ мінімізує хроматичні спотворення в 3×3 околиці біля p. $A_p(f_i, b_j)$ встановлює, що обчислені значення альфа-матеріалу корелюють з ймовірністю пікселя p, що належить передньому плану. $D_p(f_i)$ та $D_p(b_i)$ дотримуються критерію просторової спорідненості: зразки фону та переднього плану мають бути якомога ближчими до p. Константи $e_{\{N,A,f,b\}}$ визначити покарання за високе значення в будь-якому з цих показників. На практиці виявлено, що значення $e_N = 3, e_A = 2, e_f = 1$ та $e_b = 4$ мають тенденцію давати хороші результати. Таким чином, найкраща пара зразків переднього плану і фону $(\hat{f}_p, \hat{b}_p)$ для пікселя p виходить шляхом оцінки $g_p(f_i, b_j)$ для всіх можливих пар зразків:

$$(\hat{\mathbf{f}}_p, \hat{\mathbf{b}}_p) = \text{argmin}_{\mathbf{f},\mathbf{b}}\, g_p(\mathbf{f}_i, \mathbf{b}_j).$$

Нехай (F_p^g, B_p^g) будуть відповідними кольорами найкращої пари $(\hat{f}_p, \hat{b}_p)$ для p, отриманого на стадії збору. Потім обчислюємо σ_f^2 та σ_b^2 як:

$$\begin{aligned} \sigma_f^2 &= \frac{1}{N}\sum_{q\in\Omega_f} \left\|C_q - F_p^g\right\|^2, \\ \sigma_b^2 &= \frac{1}{N}\sum_{q\in\Omega_b} \left\|C_q - B_p^g\right\|^2 \end{aligned}$$

Ω_f та Ω_b околиці 5×5 пікселів з центром $(\hat{f}_p, \hat{b}_p)$ відповідно і N = 25. Значення σ_f^2 та σ_b^2 виміряти місцеві колірні варіації в околицях Ω_f та Ω_b припускаючи невеликі одновимірні гауссові кольорові розподіли навколо $(\hat{f}_p, \hat{b}_p)$. Будемо використовувати σ_f^2 та σ_b^2 на етапі уточнення зразка. Отже, вихід етапу уточнення зразка - це кортеж $\tau_p^g = (F_p^g, B_p^g, \sigma_f^2, \sigma_b^2)$ для кожного пікселя $p \in T_u$.

Розділ 3. Модифікований метод віднімання фону

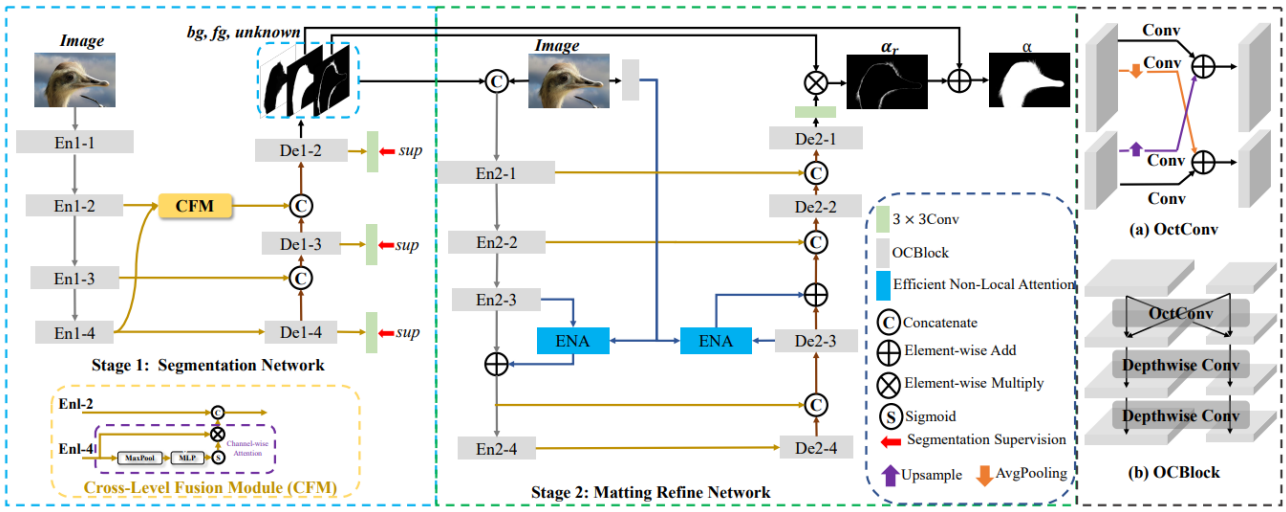
3.1 Основна ідея методу

Існують дві точки зору до вирішення цієї проблеми: полегшена магістраль і ефективні стратегії агрегації/покращення функцій на кожному етапі. Спочатку будується легкий базовий блок з деякими популярними ефективними згортковими операціями. Тоді ці базові блоки складаються, щоб побудувати легкі U-Net магістралі для двох етапів. Перший етап - це сегментація мережі (SN), а другий етап - матування уточнення мережі (MRN). На першому етапі проектується міжрівневий модуль злиття (CFM), щоб безпосередньо з'єднати високорівневу семантичну функцію з низькорівневою ознакою, яка може покращити розрізнення представлення ознак. На другому етапі проектується модуль ефективної нелокальної уваги (ENA) для моделювання відповідності між різними пікселями у функції кодера/декодера і функція зображення, витягнута з блоку (OCBlock), яка може допомогти регресувати високоякісні значення альфа. Більше того, в задачі матування, моделювання релевантності різних пікселів на дальній дистанції не так важливо, як моделювання релевантності пікселів на ближній. Тому запропонована ENA спочатку розраховує довгу релевантність за допомогою стратегії вибірки, а потім безпосередньо обчислює коротку релевантність. Така конструкція відповідає принципу легкої моделі.

SN має просту архітектуру U-Net, подібну до кодера-декодера, і магістраль утворена за допомогою стекування OCBlock. Запропонована схема має чотири рівні, іменовані як En1-1, En1-2, En1-3, En1-4. Кількість OCBlock на кожному рівні становить: 3, 4, 6, 4. Оскільки En1-1 знаходиться надто близько до входу, а його роздільна здатність занадто висока, щоб заощадити обчислювальні витрати, використовуються лише функції

останніх трьох рівнів для наступного процесу. Міжрівневий модуль злиття (CFM) додається між En1-4 і En1-2 для покращення розрізнення уявлень ознак. SN має 3 декодери, і кожен декодер складається з одного OCBlock.

Декодер об'єднує оброблені функції з кодерів і функції з підвищеною дискретизацією попереднього рівня декодера знизу вгору. Вихід кожного декодера визначається як De1-4, De1-3 і De1-2. Нарешті, мережа генерує логіти 3-канальної класифікації T за згорткою 3×3. MRN має базову архітектуру кодер-декодер, і вхід MRN є передбаченими логітами T і зображення. Основу також поєднує OCBlock і весь кодер рівні містять 2 OCB блоки відповідно. Для точної регресії альфа-значень у невідомому регіонах добавляються два ефективних модуля нелокальної уваги (ENA) до кодера та декодера симетрично в стеблі. Крім того, функції зображення, витягнуті з OCBlocks, містять більш детальну інформацію про текстуру, яка може допомогти регресувати альфа-значення. Таким чином, використовуються функції зображення, витягнуті з OCBlocks, щоб направляти потік інформації на кодері/декодері особливості. MRN має 4 декодери, і кожен декодер складається з 1 OCBlock. Запобіжники декодера функції в кожному кодері та функції з підвищеною дискретизацією з попереднього рівня знизу вгору. Нарешті, разом з невідомими картами та картами переднього плану, передбаченими SN, MRN може регресувати 1-канальний результат α.



3.2 Механізм сегментації

SN відіграє роль захоплення достатньої семантики та класифікації пікселів на невідомі, області переднього плану та заднього фону. Таким чином, вивчення різних представлень ознак з меншими обчислювальними витратами є важливим. Щоб досягти цієї мети, спочатку використовується полегшена магістраль для отримання багатомасштабних представлень функцій. Крім того, використовується модуль CFM, щоб використовувати переваги функцій в високого рівня, щоб дізнатися більше дискримінантних представлень ознак. Функція високого рівня з En1-4 містить багато семантичної інформації, яка може допомогти мережі краще розрізняти різницю між зовнішнім виглядом невідомого регіону, переднього та заднього плану. Проте SN є U-подібною архітектурою, коли інформація поступово повертається з високого рівня на нижчий, семантична інформація високого рівня поступово розбавляється. Таким чином, об'єднується низькорівнева функція En1-2 з функцією високого рівня En1-4, щоб поширити семантичну інформацію високого рівня до низького рівня. Щоб полегшити втручання нерелевантної семантичної інформації, використовується увага на каналі En1-4, щоб призначити більші ваги каналам, які важливіші для завдання матування. Операції

з увагою до каналів мають низьку обчислювальну складність, тому це вимагає незначних обчислювальних витрат.

3.3 Механізм Віднімання

MRN спрямований на регресію альфа-значень пікселів у невідомих регіонах. MRN приймає конкатенацію 3-канальних зображень і 3-канальної сегментації що є результатом SN як 6-канальний вхід. Для зручності запам'ятовування ми назвали невідомі регіони U. Проектується ефективний нелокальний модуль уваги, який може допомогти краще регресувати значення альфа з меншою кількістю обчислювальної вартості.

3.3.1 Ефективна нелокальна увага

Моделювання релевантності різних пікселів може допомогти регресувати точне значення альфа. Тому, що це може допомогти механізму ефективно групувати пікселі з подібною текстурою та звертати більше уваги на ці пікселі. Для досягнення цієї мети використовується нелокальна увага. Загальну нелокальну увагу можна визначити як:

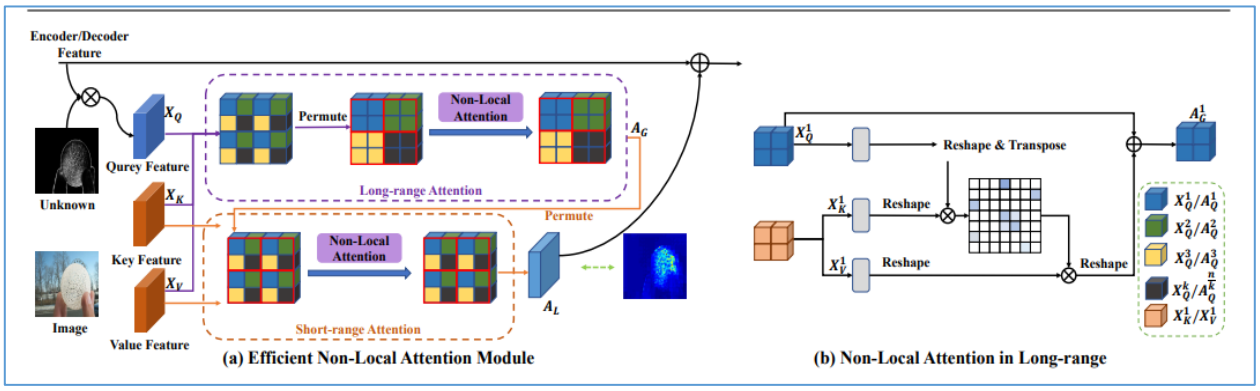
$$A = Sim(Q, K)V,$$

де Q, K і V означають запит, ключ і значення, а Sim — подібність між запитом і ключовою особливістю.

Згідно з традиційними методами, пікселі з майже ідентичною текстурою повинні мати схожу непрозорість. Отже, пікселі, які мають подібну інформацію про текстуру, повинні мати однакову непрозорість особливості. Однак у міру поглиблення мережевих шарів відображення особливості кожного пікселя буде розмитим, що вплине на регресію альфа-значення. Тому використовуються детальні функції зображення для керування потоком інформації щодо функцій кодера/декодера, що покращує точність регресії альфа-значень. У запропонованій роботі функцією запиту є функція від кодера/декодера,

названа як X_Q. Особливості "ключ/значення" - це особливості витянуті з вхідного зображення, названі як X_K та X_V.

3.3.2 Основна ідея MRN



Оскільки MRN дбає лише про регресію альфа-значень у невідомому регіоні, X_Q множиться на U щоб зробити функцію запиту більш зосередженою на невідомому регіоні. Це може бути сформульовано як:

$$X_Q = X_Q * U.$$

Безпосереднє використання звичайної нелокальної уваги в роботі займає багато часу, тобто не відповідає принципу конструкції легких моделей. Крім того, моделювання релевантності різних пікселів на великій відстані не так важливо, як моделювання релевантності пікселів на короткій відстані, а іноді призведе до певного шуму через нерелевантні збіги. Таким чином, ENA є поєднанням уваги на далеку дистанцію з стратегією вибірки та уваги на коротку дистанцію. Головний пункт уваги на далеку дистанцію - застосування нелокальної уваги до підмножин позицій, які мають довгий просторовий інтервал відстані, що може не тільки допомогти зменшити ймовірність нерелевантних збігів між пікселями великого радіусу дії, але також заощаджують витрати на обчислення. Вхідна особливість $X \in \{X_Q, X_K, X_V\}$ спочатку поділяється на k розділів, потім вибірка цих пікселів у різних розділах з інтервалами $\sqrt{k}$ у нові

підмножини. Отже, кожна підмножина містить $\frac{n}{k}$ пікселів, де n представляє розмір вхідних особливостей. Таким чином, нові підмножини визначаються як: $X_L = [X^1, X^2, \dots, X^k]$, де $X^k \in \{X_Q^k, X_K^k, X_V^k\}$, та розмірність $X^k - \mathbb{R}^{\frac{n}{k} \times C}$, C-кількість каналів. Потім застосовується нелокальна увага на X^k незалежно:

$$A_G^k = Sim(X_Q^k, X_K^k) X_V^k,$$

Нарешті об'єднуються всі A_G^k щоб отримати вихід A_G. Для коротко-дистанційної уваги, переставляємо A_G щоб згрупувати спочатку найближчі позиції разом, що є запитом на коротко-дистанційну увагу. Таким чином, вхід короткої уваги - $Z \in \{A_G, X_K, X_V\}$.

Аналогічно ділимо на $\frac{n}{k}$ розділи і кожен розділ містить k сусідніх позицій, визначених як: : $Z_S = \left[Z^1, Z^2, \dots, Z^{\frac{n}{k}}\right]$, де $Z^{\frac{n}{k}} \in \left\{A_G^{\frac{n}{k}}, X_K^{\frac{n}{k}}, X_V^{\frac{n}{k}}\right\}$, де розмірність $Z^{\frac{n}{k}} - \mathbb{R}^{k \times C}$.

Потім нелокальну увагу прикладається до кожного $Z^{\frac{n}{k}}$ незалежно. Відповідно, можемо отримати $A_L^{\frac{n}{k}}$ і об'єднати їх усі, щоб отримати результат A_L. Далі з'єднуємо A_L з функціями кодера/декодера, щоб отримати уточнені особливості. У моделі експериментально встановлено k = 16.

Разом з невідомими картами MRN виведе уточнені альфа-значення α_r в невідомих регіонах. Остаточний результат α можна записати так:

$$\alpha = F + \alpha_r.$$

Побудова функції втрат складається з двох частин: L_{SN} та L_{MRN}. Нехай α^{gt} буде нижнім істинним значенням α.

Прогноз SN - це 3-канальна класифікація логітів T. Далі генерується його нижнє істинне значення T^{gt}, що співвідноситься до α^{gt}, та визначається як:

$$T^{gt}(x,y) = \begin{cases} 0, & \alpha^{gt}(x,y) = 0 \\ 1, & \alpha^{gt}(x,y) = 1 \\ 2, & 0 < \alpha^{gt}(x,y) < 1 \end{cases}$$

де (x, y) означає розташування кожного пікселя на зображенні. Значення {0,1,2} означає що піксель знаходиться у фоновому, передньому або невідомій області. Softmax перехресні втрати ентропії використовується для нагляду за SN, який визначається як:

$$L_{SN} = \sum_{l}^{l=3} CE(T_l, T_{gt}).$$

Як показано на рисунку моделі, багаторівневий нагляд використовується як допоміжна втрата для полегшення достатності навчання. T_l означає результат прогнозу на l-рівні. Для контролю MRN, є зважені L_1 втрати і градієнтні втрати L_g. Зважені L_1 втрати визначаються як:

$$L_1 = \sum_{p} w_p \cdot |\alpha_p - \alpha_p^{gt}|,$$

Де α_p означає прогнозоване значення альфа в пікселі p. Вага w_p дорівнює 1, коли $0 < \alpha_p < 1$ і дорівнює 0.1, коли $\alpha_p = 1$ або $\alpha_p = 0$. Ця операція може змусити мережу звернути увагу на невідомий регіон. Градієнтна втрата L_g g використовується для видалення надмірно розмитого альфа, який визначається як:

$$L_g = |\nabla_x(\alpha) - \nabla_x(\alpha^{gt})| + |\nabla_y(\alpha) - \nabla_y(\alpha^{gt})|.$$

Таким чином, L_{MRN} може бути представлений як:

$$L_{MRN} = L_1 + L_g.$$

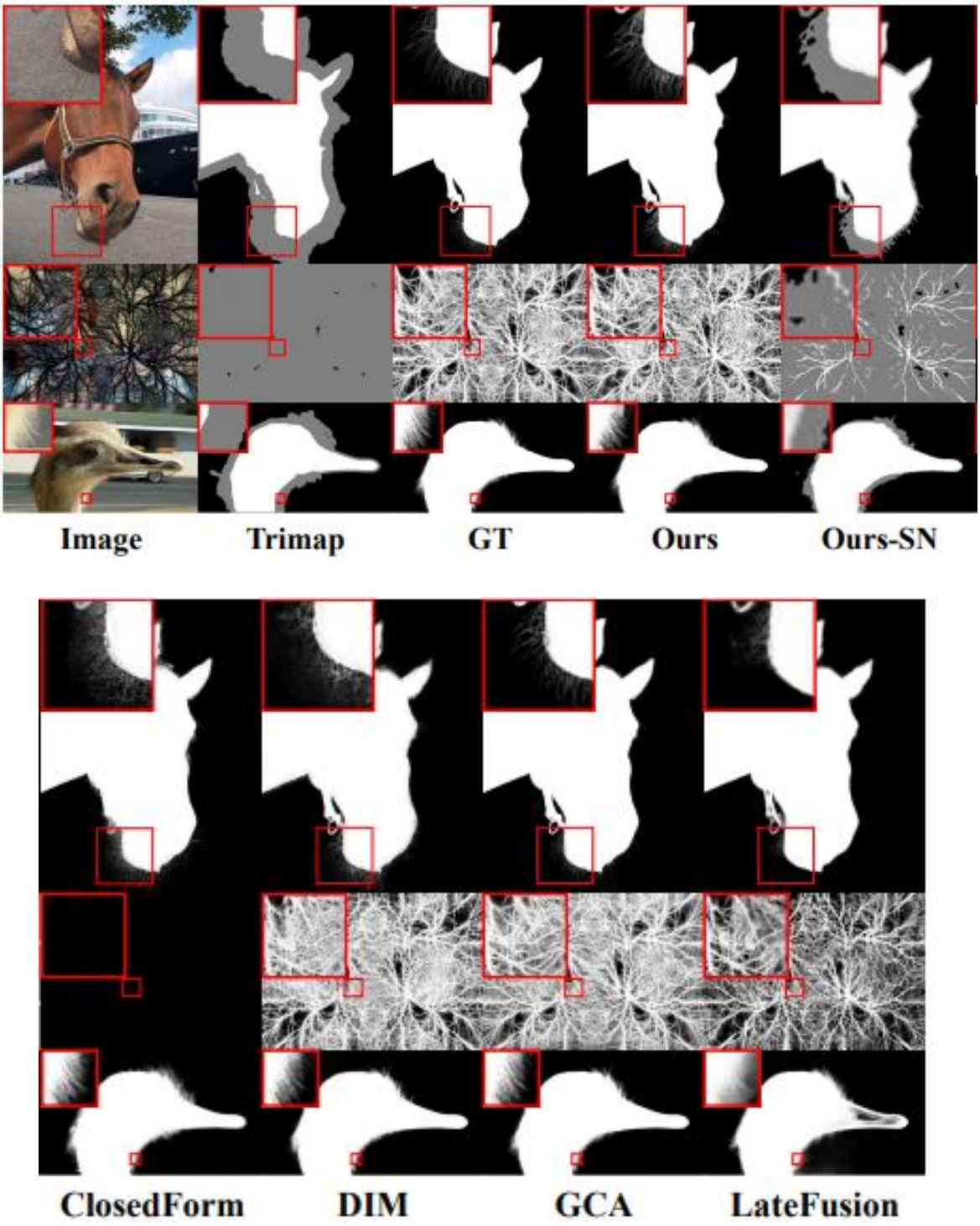
Зверніть увагу, що всі частини мережі навчаються спільно, тому загальна функція втрат задається так:

$$L = L_{SN} + L_{MRN}.$$

Розділ 4. Порівняльний аналіз модифікованого алгоритму з існуючими

4.1 Вхідні дані

Оцінка мережи відбувається на загальнодоступному Adobe Composition-1k. Навчальний набір складається з 431 об'єкта переднього плану з відповідною нижньою істиною альфа. Кожне зображення переднього плану поєднується зі 100 фоновими зображеннями від MS COCO набір даних для компонування вхідних зображень. Для тестового набору Композиція-1k містить 50 зображень переднього плану, а також відповідні альфа-матове зображення та 1000 фонових зображень з набору даних PASCAL VOC2012. Для збільшення даних, об’єкт переднього плану та альфа-зображення буде змінено до 640 × 640 зображень з імовірністю 0,25, щоб мережа побачила ціле зображення на передньому плані замість обрізаного фрагмента. Потім використовується випадкове афінне перетворення. Згодом випадковим чином зображення обрізаються по одному лоскутку 512 × 512 з кожного переднього плану. Всі лоскутки зосереджені на невідомій області відповідно до альфа-зображень. Зображення переднього плану потім перетворюються в простір HSV, а до відтінку накладаються різні тремтіння, насиченість і цінність. Під час тестування роздільна здатність вхідного зображення становить 512×512.



Візуальні порівняння в наборі даних тестування Composition-1k. Масштаб для наглядності.

Метрики

У оцінках використовуються чотири показники: SAD (сума абсолютних різностей), MSE (середня квадратична помилка), градієнт і зв'язність. Чим нижчі значення показників, тим краще прогнозований альфа.

4.2 Порівняння

Тут мережа порівнюється з 5 традиційними методами матування: Shared Matting, Learning Based, Global Matting, ClosedForm, KNN Matting та 7 методів на основі глибинного навчання: DCNN, DIM, AlphaGAN, IndexNet, Late Fusion, HAttMatting, GCA. LateFusion, HAttMatting і запропонована мережа – це методи trimap-free. Для інших методів надається RGB-зображення та трімапи випадково розширені на 25 пікселів. Обчислюється метрика 800 × 800 для всіх методів. Налаштування тестування в роботі виконується за допомогою двох trimap-free методів.

Таблиця 1.

Model	Complexity		Composition-1k			
	#PARM.	FLOPs	SAD	MSE	Grad	Conn
Shared Matting	-	-	115.20	0.031	139.88	121.35
Learning Based	-	-	100.51	0.027	107.83	104.74
Global Matting	-	-	121.46	0.034	125.11	133.23
Closed-form	-	-	118.98	0.032	130.77	118.17
KNN Matting	-	-	133.99	0.051	140.29	134.03
DCNN	-	-	122.40	0.030	129.57	121.80
DIM-Trimap-less	130.55M	443.56G	103.07	0.108	73.76	106.64
DIM	130.55M	443.56G	33.64	0.008	30.23	31.92
IndexNet	5.72M	71.41G	14.97	0.006	12.02	14.58
AlphaGAN	40.3M	496.36G	20.28	0.010	21.11	20.46
GCA	25.16M	89.41G	13.54	0.006	11.39	12.96
Late Fusion	37.91M	74.85G	58.29	0.011	36.58	59.63
HAttMatting	59.2M	159.7G	44.01	0.007	29.26	46.41
Ours	344.82K	0.93G	**19.59**	**0.006**	**18.70**	**20.34**

Підкреслення вказує на найкращий із методів на основі trimap. Цифри, виділені жирним шрифтом, представляють найкращу продуктивність у trimap-free методах.

Як видно з табл.1, метод перевершує всі традиційні методи матування зображення з великим відривом. У порівнянні з методами глибокого навчання на основі trimap, метод трохи поступається GCA і IndexNet, і краще, ніж DIM і AlphaGAN. Однак трімапи є необхідними під час їх навчання і фазі інтерференції, яка обмежує їх ефективність у практичному застосуванні. Порівняно з trimap-free методами LateFusion та HAttMatting, метод вже може перевершити ці два методи. Крім того, метод досягає продуктивності SOTA з меншою кількістю параметрів (0,3% ~ 7,8% великих моделей) і FLOP (0,2% ~ 1% великих моделей), ніж усі з перерахованих вище методів, які підвищать високу практичність.

Порівняно з LateFusion, метод здатний знайти більш повний передній план об’єктів, оскільки мережа може використовувати багатомасштабні функції для кращого захоплення семантики. Завдяки запропонованому ENA метод може регресувати більш точне значення альфа в невідомому регіоні.

4.3 Навчання

Основна ідея методу містить три частини:

- light-weighted backbone,
- cross-level fusion module (CFM)
- Efficient Non-Local Attention (ENA)

Щоб підтвердити ефективність light-weighted backbone, замінимо її іншими light-weighted backbone’ами, включаючи MobileNet і ShuffleNet і порівняємо. У порівнянні з цими двома light-weighted backbone’ами, метод light-weighted backbone може досягти кращої продуктивності з меншою кількістю параметрів і FLOP’сів. Тому що backbone може покращити багатомасштабне представлення функцій для захоплення достатньої семантики. Більше того, лише використання CFM або ENA вже може значно покращити продуктивність. Краща продуктивність була досягнута завдяки комбінації цих двох архітектур.

Backbone	Configurations		Model		Metrics			
	CFM	ENA	# Param.	FLOPs	SAD	MSE	Grad	Conn
Ours	×	×	285.70K	0.63G	56.86	0.019	36.81	48.16
	×	✓	342.70K	0.93G	34.33	0.021	28.24	32.98
	✓	×	287.83K	0.63G	28.90	0.013	24.27	26.35
	✓	✓	344.93K	0.93G	**19.59**	**0.006**	**18.70**	**20.34**
MobileNetV2	✓	✓	37.76M	27.02G	47.21	0.014	30.28	36.46
ShuffleNetV2	✓	✓	30.22M	20.52G	48.72	0.015	33.84	38.21

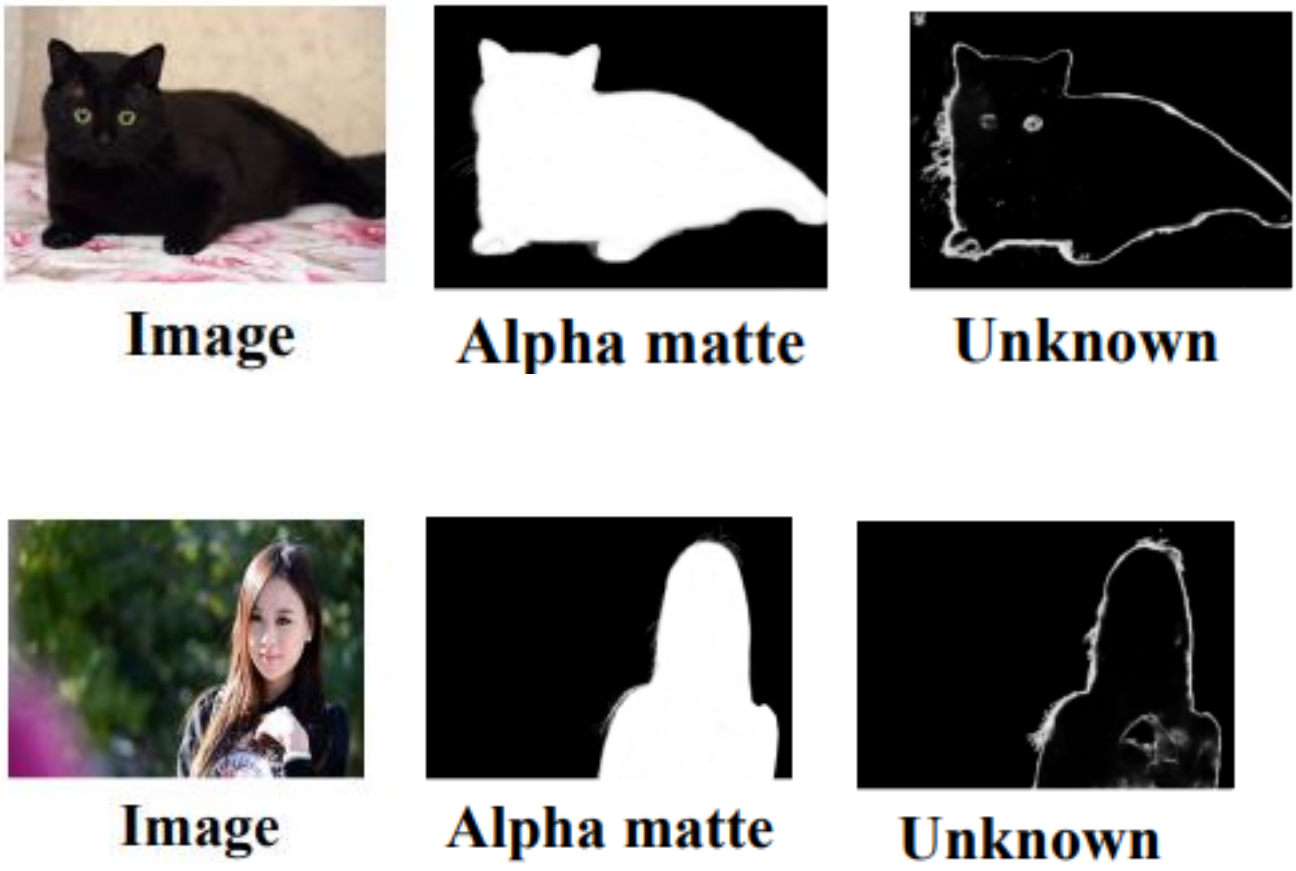
Навчання з CFM та ENA. також backbone замінено на існуючий light-weighted backbone.

4.4 Тестування моделі на реальних зображеннях

Потрібно додатково перевірити дійсність і ефективність розробленого ефективного нелокального модуля уваги. Тренування моделі проходило з різними типами уваги: звичайною нелокальною увагою та ефективною нелокальною моделлю уваги. Як показано в таблиці нижче, у порівнянні із загальною місцевою увагою, ENA є не тільки ефективнішим, але й більш придатним для віднімання фону. Використання низько-дистанційної уваги (S) далеко-дистанційної уваги (L) може покращити продуктивність моделі, оскільки вони можуть відображати релевантність між різними пікселями, що може допомогти регресу значень альфа. Як видно з таблиці, різні інтервали вибірки $\sqrt{k}$ призводять до різних результатів. Коли $\sqrt{k} = 1, 2$ ENA не може повністю проявити свою зручність та ефективність, тому що вибірка надто інтенсивна. І далеко-дистанційна увага призведе до деяких артефактів через нерелевантні збіги. Експериментально встановлено, що $\sqrt{k} = 4$ дає кращі результати у всіх експериментах.

Configurations		Model		Metrics			
		# Param.	FLOPs	SAD	MSE	Grad	Conn
Common Non-Local		421.69K	1.33G	25.13	0.010	20.89	22.67
Efficient Non-local(Ours)	S	325.15K	0.82G	25.98	0.011	21.93	23.02
	L($\sqrt{k}$=4)	292.19K	0.72G	27.01	0.013	23.94	24.31
	S+L($\sqrt{k}$=1)	424.11K	1.34G	25.01	0.010	20.31	22.96
	S+L($\sqrt{k}$=2)	371.43K	1.02G	23.27	0.009	19.37	21.61
	S+L($\sqrt{k}$=4)	344.83K	0.93G	**19.59**	**0.006**	**18.70**	**20.34**
	S+L($\sqrt{k}$=8)	340.27K	0.91G	19.90	0.008	19.05	21.35

Навчання різних модальностей, використовуючи увагу. S і L означає низько-дистанційну і далеко-дистанційної увагу відповідно.



Результати на зображеннях реального світу, у тому числі невідомий регіон, передбачений методом

Висновки

В роботі розроблена двоступенева light-weighted trimap-free модель віднімання фону зображення, у сфері що активно розвивається. SN призначений для захоплення достатньої кількості семантики та класифікування пікселів на невідому область, передній план та фон. MRN спрямований на отримання детальної інформації про текстуру та регресію точних альфа-значень. З запропонованих CFM, SN може ефективно використовувати багатомасштабні функції з меншими витратами на обчислення. ENA в MRN може ефективно моделювати схожість між різними пікселями та допомагати регресувати високоякісні значення альфа. Результати експерименту на тестовому наборі Composition-1k демонструють що метод перевершує найсучасніші (SOTA) підходи до віднімання фону з параметрами 1% (344k) великих моделей. Вважається, що пропонована backbone, CFM і ENA може сприяти вирішенню інших завдань, наприклад, комп'ютерного зору в майбутньому. Проблема віднімання фону є та буде дуже актуальною, бо використовується у багатьох сферах: від медицини і комп'ютерних наук, до комп'ютерної графіки та фотомистецтва. У майбутньому метод допоможе у вирішенні цієї проблеми у великих масштабах завдяки своїй оптимізації та якісним показникам.

Список Літератури

1. Yagiz Aksoy, Tunç Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In CVPR, 2017.
2. Shaofan Cai, Xiaoshuai Zhang, and Haoqiang Fan. Disentangled image matting. In ICCV, 2019
3. Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. KNN matting. In CVPR, 2012
4. Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In ACM MM, 2018.
5. Donghyeon Cho, Yu-Wing Tai, and In-So Kweon. Natural image matting using deep convolutional neural networks. In ECCV, 2016.
6. François Chollet. Xception: Deep learning with depthwise separable convolutions. In CVPR, 2017
7. Yung-Yu Chuang, Brian Curless, David Salesin, and Richard Szeliski. A bayesian approach to digital matting. In CVPR, 2001.
8. Eduardo Simoes Lopes Gastal and Manuel M. Oliveira. Shared sampling for real-time alpha matting. CGF, 29(2), 2010.
9. Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In CVPR, 2011.
10. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.
11. A. Al-Kabbany and E. Dubois. Matting with sequential pair selection using graph transduction. In 21st International Symposium on Vision, Modeling, and Visualization, 2016.
12. V. Badrinarayanan, A. Handa, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293, 2015.

13. A. Berman, A. Dadourian, and P. Vlahos. Method for removing from an image the background surrounding a selected object, Oct. 17 2000. US Patent 6,134,346.
14. Q. Chen, D. Li, and C.-K. Tang. Knn matting. IEEE transactions on pattern analysis and machine intelligence, 35(9):2175–2188, 2013.
15. B. He, G. Wang, C. Shi, X. Yin, B. Liu, and X. Lin. Iterative transductive learning for alpha matting. In 2013 IEEE International Conference on Image Processing, pages 4282– 4286. IEEE, 2013.
16. BAI X., SAPIRO G.: A geodesic framework for fast interactive image and video segmentation and matting. ICCV 1 (2007).
17. CRABB R., TRACEY C., PURANIK A., DAVIS J., SANTA CRUZ C., CANESTA I., SUNNYVALE C.: Real-time foreground segmentation via range and color imaging. In CVPR Workshop on Time-of-flight Computer Vision (2008)
18. RHEMANN C., ROTHER C., GELAUTZ M.: Improving color modeling for alpha matting. In BMVC (2008)
19. RUZON M., TOMASI C.: Alpha estimation in natural images. In CVPR (2000), vol. 1
20. Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In ICCV, 2009.