Colorizer

Бабанский Виталий, Бакин Денис

21 марта 2025 г.

1 Введение

Задача восстановления цветных изображений из черно-белых снимков является распространенной задачи и применяется, например, при обновлении исторических снимков, которые были сделаны до изобретения цветной фотографии. Более сложной постановкой той же задачи считается раскрашивание снимков NIR (near-infrared spectroscopy) — это снимки, где вместо количества видимого света фотосенсором камеры подсчитывается количестве фотонов с длиной волны от 780 нм до 2500 нм, то есть выше видимого диапазона. Такая съемка применяется при низкой освещенности и при съемке архитектурных объектов.

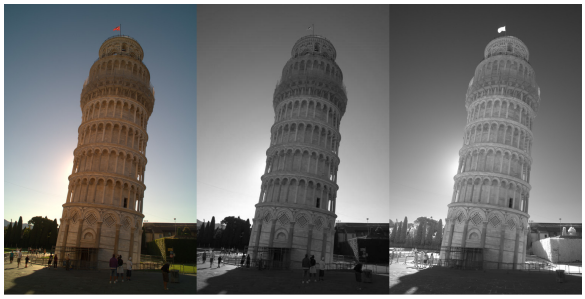

Рис. 1: Три пространства цветов: RGB, черно-белое и NIR

2 Постановка задачи

Целью проекта является создание и обучение нейронной сети для получения цветных изображений по данным черно-белым изображениям, а также провести ряд экспериментов, воспроизвести результаты выбранных статей и измерить полученное качество по набору метрик.

3 Литература

Список рассмотренных не окончательный. Включены только те статьи, идеи которых скорее всего будут использованы в реализации.

3.1 Раскрашивание с подсказками

(ZHANG; ZHU и др., 2017) предлагает архитектуру полносверточной нейронной сети, которая принимает на вход черно-белое изображение и набор локальных и глобальных подсказок от пользователя. Сеть раскрашивает указанные пиксели так, как скажет пользователь, а остальное изображение так, чтобы оно было наиболее естественным. Сеть показывает приемлемое качество как бейзлайн: основная полносверточная нейросеть используется в некоторых других более сложных архитектурах. Отличная качество достигается в особо сложных случаях,

когда на снимке есть мелкий орнамент или цвета, которые сложно восстановить из контекста (воздушные шары, например).

3.2 Instance colorization

(SU; CHU; HUANG, 2020) использует основную полносверточную нейронную сеть, из (ZHANG; ZHU и др., 2017). Идея авторов заключается в генерации ограничивающих прямоугольников (bounding boxes) вокруг известных объектов на изображении с помощью предтренированного детектора (HE и др., 2018). Затем с помощью выбранного backbone раскрашиваются как вырезанные объекты, так и все изображение в целом. Затем на этапе карт признаков модуль слияния "мягко"объединяет вырезанные раскрашенные объекты и полное раскрашенное изображение. Это дает улучшенные результаты по сравнению с прошлыми статьями и относится к полностью автоматическому раскрашиванию черно-белых снимков.

3.3 Cooperative colorization

(YANG; CHEN; YANG, 2023) авторы решают целых 2 проблемы: улучшают качество раскрашивания и предлагаются объединение и трансфер знаний модели между двумя доменами входных данных: черно-белых и NIR изображений. В статье предлагается генерировать альтернативный домен по данному (NIR по черно-белому изображению или наоборот). Затем каждое из изображений раскрашивается, результат объединяется. Поскольку такое количество генеративных сетей может отклоняться от ответа, авторы статьи предлагают множество дополнительных ограничений для модели в виде функций потерь, которые требуют, чтобы раскрашенные изображения из разных доменов были очень похожи по структуре (ведь цвет ее не меняется).

3.4 NIR-to-RGB Spectral Translation with Mamba

(ZHAI и др., 2024) является лучшей на данный момент архитектурой для раскрашивания NIR изображений (на датасете NIR изображений с подготовленной валидационной выборкой (JIE, 2020)). Основа подхода заключается в построении двух наборов связанных модулей: сети для раскрашивания в пространство RGB, сети для раскрашивания в пространство HSV, а также набора более компактных неглубоких подмодулей, описанных авторами статьи.

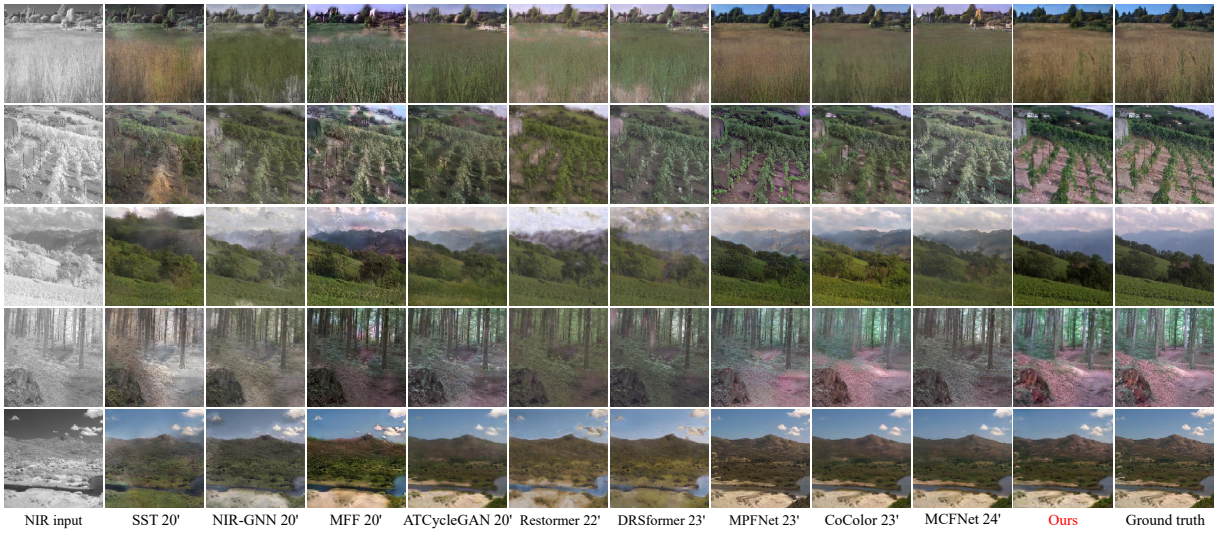

Рис. 2: Результаты работы ColorMambda (ZHAI и др., 2024). Также приведено сравнение с CoColor (YANG; CHEN; YANG, 2023)

4 Данные

Авторы статей (ZHANG; ZHU и др., 2017), (SU; CHU; HUANG, 2020), (YANG; CHEN; YANG, 2023) использовали датасеты COCO (**COCO**) и ImageNet (**ImageNet**). Авторы статьи (ZHAI и др., 2024) использовали датасет NIR изображений с подготовленной валидационной выборкой (JIE, 2020). Мы планируем использовать выборки из датасетов ImageNet, который содержит сфокусированные фотографии различных объектов, и COCO, который содержит более общие сцены: архитектуры, природы.

Возможно, будут проведены эксперименты с созданием генерации черно-белого изображения по NIR данным. В этом случае к данным будет добавлен датасет "RGB-NIR Scene Dataset".

5 Метрики качества

Для оценки качества раскрашивания снимков будем использовать набор метрик. По ним же будем сравнивать качество работы моделей.

- **MSE**. Один из наиболее очевидных методов оценки близости предсказания к "верному"ответу. К недостаткам этой метрики можно отнести неразличимость мелкой зашумленности и отсутствия контроля за резкими переходами цветов, которые требуются при корректном раскрашивании изображений.

$$MSE(I_1, I_2) = \sum_{x=1}^{W} \sum_{y=1}^{H} \frac{(I_1(x,y) - I_2(x,y))^2}{W \cdot H}$$

- **PSNR (Пиковое отношение сигнал/шум)**: PSNR измеряет качество цветных изображений, сравнивая пиксельные различия между оригиналом и раскрашенным изображением. Более высокие значения PSNR указывают на лучшее качество изображения с меньшими искажениями.

$$PSNR(I_1, I_2) = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE(I_1, I_2)} \right)$$

  где MAX — максимальное значение пикселя (например, 255 для 8-битных изображений).

- **SSIM (Индекс структурного сходства)**: SSIM оценивает структурное сходство между оригиналом и раскрашенным изображением, учитывая яркость, контрастность и текстуру. Этот индекс предоставляет более точную для восприятия меру качества изображения по сравнению с метриками на основе пикселей, такими как PSNR.

$$SSIM(I_1, I_2) = \frac{(2\mu_1\mu_2 + C_1)(2\sigma_{12} + C_2)}{(\mu_1^2 + \mu_2^2 + C_1)(\sigma_1^2 + \sigma_2^2 + C_2)}$$

  где μ_1 и μ_2 — средние значения яркости оригинала и раскрашенного изображения, σ_1^2 и σ_2^2 — дисперсии, σ_{12} — ковариация, а C_1 и C_2 — константы для стабилизации деления. По приведенной формуле метрика считается локально, а затем усредняется по всему изображению.

- **AE (Абсолютная ошибка)**: AE количественно оценивает абсолютное различие между соответствующими пикселями оригинала и раскрашенного изображения. Меньшие значения AE указывают на лучшую точность раскраски.

$$AE(I_1, I_2) = \sum_{x=1}^{W} \sum_{y=1}^{H} \frac{|I_1(x,y) - I_2(x,y)|}{W \cdot H}$$

- **LPIPS (Обученное перцептуальное сходство изображений)** (ZHANG; ISOLA и др., 2018): LPIPS оценивает перцептуальное сходство с использованием моделей глубокого обучения, фокусируясь на том, как человеческое зрение воспринимает различия между оригиналом и раскрашенным изображением. Более низкие значения LPIPS означают, что раскрашенным изображение более точно соответствует восприятию человека.

$$LPIPS(I_1, I_2) = \frac{1}{N} \sum_{i=1}^{N} \|f_i(I_1) - f_i(I_2)\|_2^2$$

где f_i — выходные данные i-го слоя предобученной модели, а N — количество слоев.

6 Текущая идея

Идея на момент написания отчета и вероятно изменится в будущем.

В качестве бейзлайна хочется реализовать полносверточную нейронную сеть из (ZHANG; ZHU и др., 2017) и провести ряд экспериментов, возможно, с реализацией интерактивных подсказок в пользовательском интерфейсе. Затем хотелось бы реализовать одну из других рассмотренных статей и проверить воспроизводимость результатов по выбранным метрикам.

7 Бейзлайн

В качестве бейзлайна согласно (ZHANG; ZHU и др., 2017) была выбрана полносверточная нейронная сеть — UNet, которая изначально была предложена для сегментации медицинских изображений, (RONNEBERGER; FISCHER; BROX, 2015). На вход такая сеть получает одноканальное изображение (более подробно преобразования изображений будут описаны в следующем разделе), затем применяет набор сверток с residual connections, сокращая в текущей модификации размер изображений в 8 раз по каждому измерению, после чего upscale свертками восстанавливает его исходные размер.

Выход сети — двуканальное изображение, которое в изначальной статье интерпретировалось как распределение по двум классам каждого из пикселей: наибольшая вероятность у того класса маски сегментации, к которому скорее всего приналежит текущий пиксель. В нашем бейзлайне двуканальный выход сети интерпретировался как два нормированных цветовых канала.

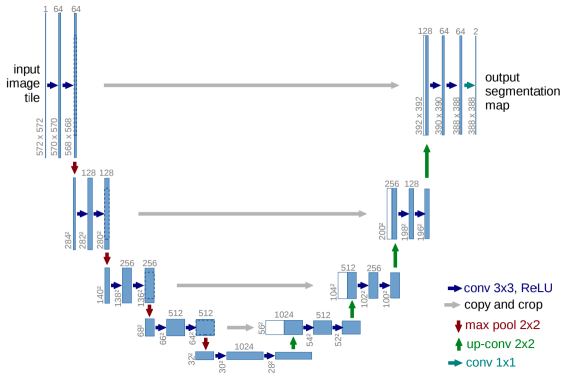

Рис. 3: UNet архитекрута (пример для 32x32 в середине сети). Каждый голубой прямоугольник соответствует многоканальной карте признаков

8 Обучающий пайплайн

Опишем, какие преобразование применяются к изображениям.

1. Читаем RGB изображение
2. Преобразуем RGB изображение в цветовое пространство CIELab и разделяем на одиночный канал L – яркость пикселей, вход сети и и на двуканальную матрицу ab
3. нормируем L и ab на максимальные значение по каналам: 100 и 255 соответственно
4. подаем L на вход сети, получаем ab
5. денормируем каналы
6. совмещаем в трехкальное CIELab изображение и преобразуем обратно в RGB
7. показываем пользователю или сохраняем

Для удобства разработки и постановки экспериментов используется создание и сохранение логов с помощью библиоткеи Wandb, в репозитории сохраняется модульная архитектура Python файлов, почти везде выбран объектно-ориентированный подход. Например, в классе Trainer одноименного модуля реализована вся логика, связанная с обучением, дообучением, сохранением, загрузкой, тестированием моделей с выбранной функцией потерь, оптимизатором и загрузчиком данных для обучения, валидации и тестирования.

Ссылка на репозиторий: https://github.com/dfbakin/colorizer/tree/checkpoint-1-baseline

Список литературы

HE, Kaiming и др. **Mask R-CNN**. [**sinelocosinenomine**], 2018. arXiv: 1703.06870 [cs.CV]. Режим доступа: <https://arxiv.org/abs/1703.06870>.

JIE, Chen. **VCIP 2020 Grand Challenge on NIR Image Colorization**. [**sinelocosinenomine**], 2020. Режим доступа: <https://jchenhkg.github.io/projects/NIR2RGB_VCIP_Challenge/>.

RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. [**sinelocosinenomine**], 2015. arXiv: 1505.04597 [cs.CV]. Режим доступа: <https://arxiv.org/abs/1505.04597>.

SU, Jheng-Wei; CHU, Hung-Kuo; HUANG, Jia-Bin. **Instance-aware Image Colorization**. [**sinelocosinenomine**], 2020. arXiv: 2005.10825 [cs.CV]. Режим доступа: <https://arxiv.org/abs/2005.10825>.

YANG, Xingxing; CHEN, Jie; YANG, Zaifeng. Cooperative Colorization: Exploring Latent Cross-Domain Priors for NIR Image Spectrum Translation. B: PROCEEDINGS of the 31st ACM International Conference on Multimedia. [**sineloco**]: ACM, окт. 2023. (MM '23), с. 2409—2417. DOI: 10.1145/3581783.3612008. Режим доступа: <https://arxiv.org/abs/2308.03348>.

ZHAI, Huiyu и др. **ColorMamba: Towards High-quality NIR-to-RGB Spectral Translation with Mamba**. [**sinelocosinenomine**], 2024. arXiv: 2408.08087 [cs.CV]. Режим доступа: <https://arxiv.org/abs/2408.08087>.

ZHANG, Richard; ISOLA, Phillip и др. **The Unreasonable Effectiveness of Deep Features as a Perceptual Metric**. [**sinelocosinenomine**], 2018. arXiv: 1801.03924 [cs.CV]. Режим доступа: <https://arxiv.org/abs/1801.03924>.

ZHANG, Richard; ZHU, Jun-Yan и др. **Real-Time User-Guided Image Colorization with Learned Deep Priors**. [**sinelocosinenomine**], 2017. arXiv: 1705.02999 [cs.CV]. Режим доступа: <https://arxiv.org/abs/1705.02999>.