УДК 004.8:004.4

СЕГМЕНТАЦИЯ ИЗОБРАЖЕНИЙ ГОЛОВЫ ЧЕЛОВЕКА НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ

М. М. Лукашевич[1), 2)], В. В. Венгеренко[2)], А. А. Воронов[2)]

[1)] *Белорусский государственный университет,*
Минск, Беларусь, lukashevichmm@bsu.by
[2)] *Объединённый институт проблем информатики НАН Беларуси,*
Минск, Беларусь, lukashevichmm@bsu.by, vengerenko@lsi.bas-net.by,
voronov@lsi.bas-net.by

В статье рассматривает сегментации головы человека на изображениях с целью последующего построения 3D-модели для биометрической идентификации и сравнения с данными из видеонаблюдения. Представлены и протестированы современные архитектуры нейронных сетей: U-Net, PortraitNet, Segmenter и Segment Anything Model (SAM). Эксперименты проведены на наборе данных из пар «изображение–маска» с различными ракурсами.

Ключевые слова: сегментация изображений; удаление фона; нейронные сети; U-Net; Segmenter; предобработка видеоданных; алгоритмы компьютерного зрения.

HUMAN HEAD IMAGE SEGMENTATION BASED ON NEURAL NETWORKS

M. M. Lukashevich[a), b)], V. V. Vengerenko[b)], A. A. Voronov[b)]

[a)] *Belarusian State University,*
Minsk, Belarus, lukashevichmm@bsu.by
[b)] *The United Institute of Informatics Problems,*
Minsk, Belarus, lukashevichmm@bsu.by, vengerenko@lsi.bas-net.by,
voronov@lsi.bas-net.by

The article addresses human head segmentation in images for subsequent 3D model reconstruction, aimed at biometric identification and comparison with surveillance data. Modern neural network architectures (U-Net, PortraitNet, Segmenter, and Segment Anything Model (SAM)) are presented and tested. Experiments are conducted on a dataset of image-mask pairs with various viewing angles.

Keywords: image segmentation; background removal; neural networks; U-Net; Segmenter; video data preprocessing; computer vision algorithms.

1. Введение

Сравнение фотографий человека, полученных в различных условиях, с существующими фотографиями, хранящимися в базах данных, является распространенной практикой для принятия решений. Такие виды

сравнения используются в основном для идентификации личностей преступников, подозреваемых, жертв и неопознанных лиц, а также в других аналогичных областях. В ходе таких сравнений используются изображения человека в реальной сцене и изображения, сделанные в контролируемой среде. Однако углы обзора объекта интереса (человека) в подавляющем большинстве случаев не совпадают, т.к. изображения реальных сцен регистрируются случайным образом, а изображения человека в контролируемой среде делаются под прямым углом. Очевидно, что произвести идентификацию личности достаточно сложно. В этом случае требуется другой ракурс анализа изображений, а он может быть доступен. Одним из возможных решений является технология получения 3D моделей головы человека. На ее основе возможно получение изображения с заданным углом поворота головы [1].

В общепринятой методологии построения 3D модели головы человека условно выделяют несколько этапов:

1) установка параметров положения и поворота устройства ввода (камеры) является тривиальной задачей и может быть оптимизирована с помощью обычных инженерных методов;

2) получение изображений из видео и выделение силуэта человека (удаление фона) на этих изображениях;

3) сегментация изображения головы человека на изображении для лучшего выравнивания и измерения;

4) удаление неинформативных артефактов и определением очертаний объектов;

5) построение трехмерной сетки и вписывание в нее полученных изображений;

6) текстурирование.

Как указано, технология генерации компьютерной 3D модели головы человека для сравнения с изображениями, полученными с камеры, включает этап сегментации головы человека, отделение ее от фона. Указанная задача и является целью данной работы, а ее результаты станут элементом разрабатываемой технологии построения 3D модели головы.

2. Алгоритмы удаления фона на изображениях головы человека

Удаление фона на изображениях головы человека представляет собой важное направление в области компьютерного зрения и цифровой обработки изображений. Эта задача находит применение в телемедицине, системах дополненной реальности, видеоконференцсвязи, биометрической идентификации, а также в инструментах цифрового редактирования.

Наиболее заметные успехи в данной области связаны с использованием глубоких сверточных нейронных сетей. Архитектуры, такие как U-Net, DeepLabV3+ и Mask R-CNN, показывают высокое качество

сегментации благодаря своей способности моделировать сложные зависимости между пикселями. Эти модели обеспечивают детализированное выделение контуров лица и хорошо справляются с разнообразием фонов и условий освещения. Недостатком является необходимость использования значительных вычислительных мощностей и наличие большого объёма размеченных данных для обучения.

Для работы на мобильных устройствах и в реальном времени предложены оптимизированные архитектуры, такие как LiteUNet, MODNet и PortraitNet. Они сочетают в себе приемлемое качество сегментации и высокую скорость обработки, что делает их подходящими для внедрения в потребительские устройства и облачные сервисы с ограничениями по производительности.

Одним из наиболее перспективных направлений является использование Transformer-архитектур, таких как Segmenter, TransUNet и Swin-Unet, которые способны эффективно захватывать контекст и пространственные зависимости, особенно в сложных сценариях. Также набирает популярность комбинация CNN и Transformer для создания гибридных моделей, сочетающих преимущества обоих подходов.

В условиях недостатка размеченных данных всё большее значение приобретают методы самообучения и few-shot learning. Модели, такие как DINO, MAE и Segment Anything Model (SAM), открывают новые возможности для применения таких нейронных сетей без необходимости полной аннотации данных.

Исходя из результатов анализа определены следующие перспективные архитектуры для сегментации и удаления фона на изображениях головы человека: U-Net, PortraitNet, Segmenter и SAM [2–7].

3. Экспериментальные исследования

Выполнено прототипирование обучения и тестирования нейронных сетей на основе языка программирования Python, версия 3.12.11 с использованием GPU T4. Использованные при написании кода ключевые библиотеки, а также их назначение приведены в табл. 1.

Таблица 1

Ключевые библиотеки для прототипирования алгоритмов удаления фона на основе нейронных сетей

Библиотека	Назначение
torch	Создание архитектуры нейронной сети, работа с GPU, работа с тензорами, реализация алгоритма обучения, а именно вычисление функции потерь и оптимизация весов нейронной сети; загрузка набора данных, батчей, итерации по обучающим и тестовым данным. Загрузка MobileNetV2 как backbone для архитектуры PortraitNet, использование предобученных весов (IMAGENET1K_V1) для переноса обучения.

Окончание табл. 1

Библиотека	Назначение
torchvision	Изменение размера изображения, конвертация в тензор, нормализация изображения. Нормализация по ImageNet-статистикам.
Pillow	Работа с изображениями в Python, загрузка изображений, перевод изображений из RGB в полутон.
scikit-learn	Разбиение на обучающую и тестовую выборки набора изображений, вычисление метрик качества.
transformers	Загрузка модели, загрузка препроцессора, дообучение модели SegFormer B0 на пользовательских данных.
segment-anything	Загрузка архитектуры SAM, создание предиктора, сегментация по точкам
cv2	Загрузка и обработка изображений

Основой для создания моделей нейронных сетей являются наборы данных. Для обучения и тестирования моделей использовался набор данных, содержащий изображений и маски сегментации в формате .png размером 512×512 пикселей. Пример изображения и соответствующей ему маски приведены на рис. 1.

Всего набор данных содержит 730 изображений и соответствующих им масок. Из них 307 изображений головы со спины, 428 изображений – вид спереди головы человека, а также 211 изображений – вид сбоку. 584 изображения использовались для обучения нейронных сетей, а 146 изображений – для тестирования.

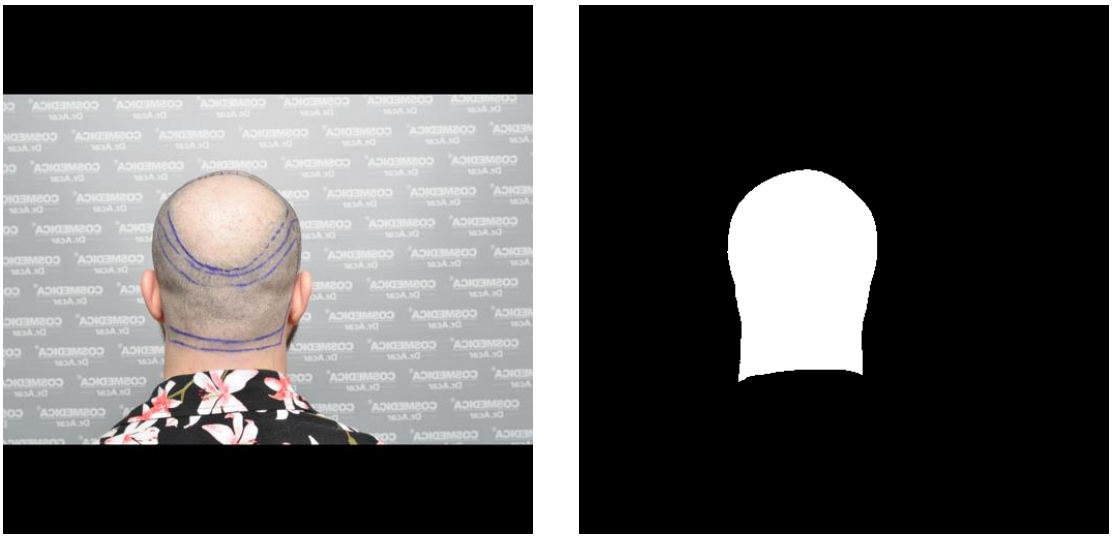

Рис. 1. Изображение головы человека и соответствующая ему маска

Сегментация головы человека фактически сводится к задаче бинарной сегментации, где присутствует фон и один класс. В этом случае

необходимо использовать следующие метрики для оценки качества сегментации [8–10].

IoU (Intersection over Union) или Jaccard Index – отношение пересечения (предсказанной и истинной масок) к их объединению:

$$IoU = \frac{TP}{TP+FP+FN}, \quad (1)$$

где *TP* (True Positive) – пиксели, правильно предсказанные как объект; FP (False Positive) – пиксели, ошибочно предсказанные как объект (ложные срабатывания); *FN* (False Negative) – пиксели, ошибочно предсказанные как фон (пропущенные объекты).

Dice коэффициент:

$$Dice = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}, \quad (2)$$

где *TN* (True Negative) – пиксели, правильно предсказанные как фон.

Accuracy (точность) – доля правильно классифицированных пикселей:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (3)$$

Каждая из трёх архитектур нейронных сетей (U-net, PortraitNet, Segmenter) была обучена на обучающем наборе данных и протестирована на тестовом наборе данных. Модель SAM (sam_vit_b_01ec64.pth) не дообучалась, а тестировалась на тех же тестовых данных. Были определены координаты референсных точек для сегментации. В табл. 2 приведены результаты тестирования обученных моделей и модели SAM. Результаты показывают, что максимальные значения трёх метрик для оценки качества сегментации показывает модель Segmenter.

Таблица 2

Результаты тестирования моделей

Архитектура	IoU	Dice	Accuracy
U-net	0,9307	0,9641	0,9927
PortraitNet	0,9159	0,9560	0,9914
Segmenter	**0,9739**	**0,9867**	**0,9973**
SAM	0,6634	0,7931	0,9597

На рис. 2–4 представлены графики обучения (изменение функции потерь) и тестирование моделей на тестовых данных.

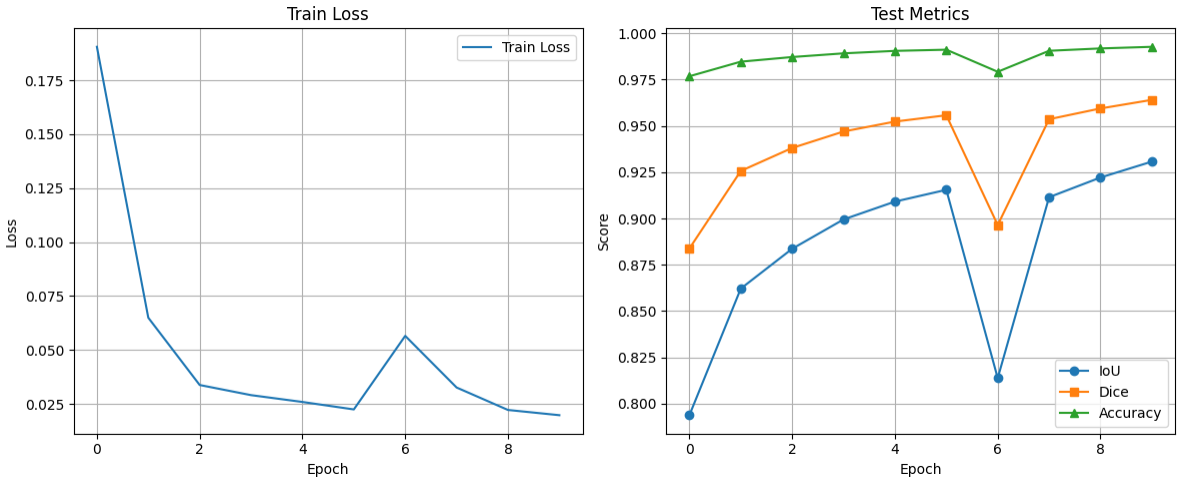

Рис. 2. Графики обучения (изменение функции потерь) U-net и тестирования модели U-net

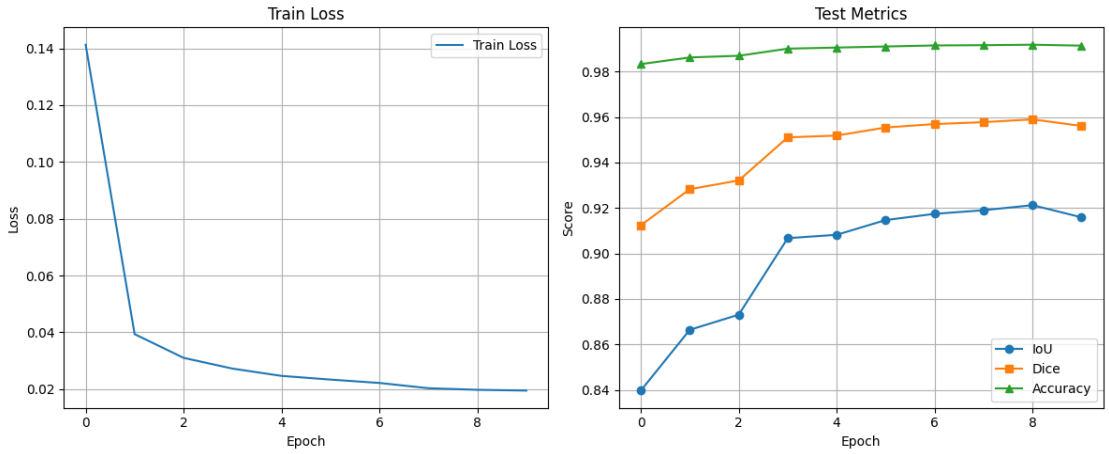

Рис. 3. Графики обучения (изменение функции потерь) и тестирования модели PortraitNet

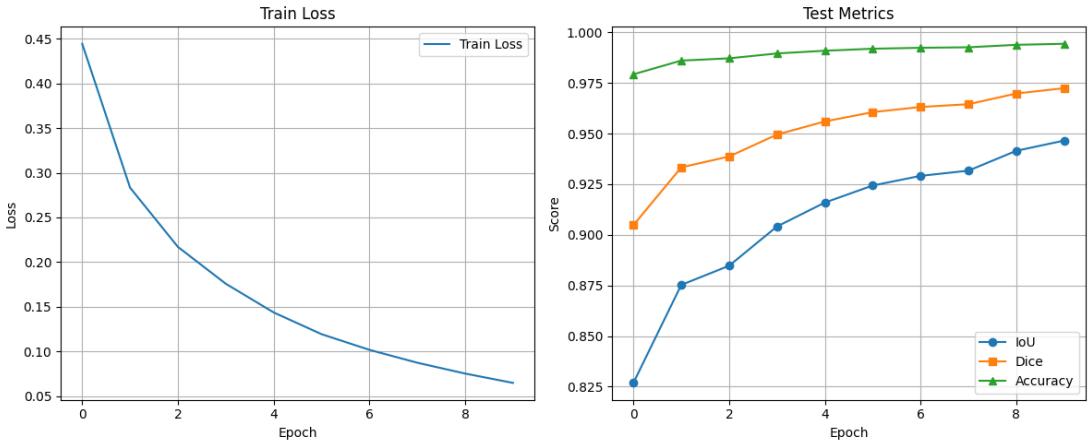

Рис. 4. Графики обучения (изменение функции потерь) и тестирования модели Segmenter

Проведенное в работе исследование выполнено в рамках совместного белорусско-турецкого проекта № Ф25ТУРГ-001.

4. Заключение

Для решения задачи удаления фона на изображениях головы человека были выбраны и протестированы четыре современные архитектуры нейронных сетей: U-Net, PortraitNet, Segmenter и SAM. Модели U-Net, PortraitNet и Segmenter были обучены на наборе данных, представленных исходными изображения головы человека в разных ракурсах и эталонными масками сегментации. Модель SAM использовалась без дообучения с предварительно расстановкой опорных точек для сегментации.

Результаты тестирования показали, что наилучшую точность демонстрирует модель Segmenter (IoU = 0,9739, Dice = 0,9867, Accuracy = = 0,9973), что делает её наиболее подходящей для задачи сегментации головы. Модели U-Net и PortraitNet также показали высокие результаты, тогда как SAM, несмотря на гибкость, показал значительно более низкое качество (IoU = 0,6634) из-за сложности автоматического определения опорных точек. Полученные результаты станут элементом разрабатываемой технологии построения 3D модели головы человека.

Библиографические ссылки

1. Towards fast, accurate and stable 3D dense face alignment / J. Guo [et al.] // Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX / Cham : Springer International Publishing, 2020. P. 152–168.

2. *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation. Freiburg : University of Freiburg, 2015.

3. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation / L.-C. Chen [et al.]. Mountain View : Google Inc., 2018.

4. Convolutional Neural Networks for Large-Scale Remote Sensing Image Classification / E. Maggiori [et al.]. Cambridge : HAL, 2017.

5. *Shelhamer E., Long J., Darrell T.* Fully Convolutional Networks for Semantic Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016. Vol. 39, № 4. P. 640–651.

6. *Badrinarayanan V., Kendall A., Cipolla R.* SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017. Vol. 39, № 12. P. 2481–2495.

7. Pyramid Scene Parsing Network / H. Zhao [et al.] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. № 1. P. 6230–6239.

8. *Szeliski R.* Computer Vision: Algorithms and Applications. 2nd ed. Cham : Springer, 2022.

9. *Goodfellow I., Bengio Y., Courville A.* Deep Learning. Cambridge : MIT Press, 2016.

10. *Prince S. J. D.* Computer Vision: Models, Learning, and Inference. Cambridge : Cambridge University Press, 2012.