



DOI: 10.15514/ISPRAS-2023-35(6)-16

Использование переноса стиля как способ улучшения обобщающей способности нейросети в задаче детекции объектов

^{1,2} Д.К. Карачёв, ORCID: 0000-0002-1008-2535 <denis.karachev@ocrv.ru>

¹ С.Е. Штехин, ORCID: 0000-0003-2866-4864 <sergei.shtekhin@ocrv.ru>

² В.С. Тарасян, ORCID: 0000-0002-5288-5221 <vtarasyan@gmail.com>

¹ И.Ю. Смолин, ORCID: 0000-0002-7272-1232 <ilya.smolin@ocrv.ru>

¹ М.В. Исаков, ORCID: 0000-0002-8677-8862 <maksim.isakov@ocrv.ru>

¹ Филиал №11 ООО «ОЦРВ» Сириус,

354340, Россия, п.г.т. Сириус, ул. Международная 2/1.

² Уральский государственный университет путей сообщения,
620034, Россия, г. Екатеринбург, ул. Колмогорова, 66.

Аннотация. В данной работе предлагается реализация подхода обучения нейронной сети для детекции объектов с помощью аугментации - переноса стиля. Данный метод улучшает обобщающую способность нейросети для определения местоположения объектов на изображении за счет улучшения взаимодействия с низкоуровневыми признаками, такими как текстуры, цвета и небольшие изменения форм. Экспериментально доказана эффективность метода и продемонстрированы числовые значения метрик детекции на нескольких наборах данных с различными классами. Применение аугментации предлагается с помощью ранее не использованной архитектуры нейросети, способной переносить произвольное количество стилей. Особенностью подхода также является то, что веса нейросети для стилизации замораживаются и она добавляется в граф детекционной сети, что позволяет увеличить скорость аугментации.

Ключевые слова: нейронные сети; компьютерное зрение; перенос стиля; машинное зрение; машинное обучение; детектирование объектов.

Для цитирования: Карачев Д.К., Штехин С.Е., Тарасян В.С., Смолин И.Ю., Исаков М.В. Использование переноса стиля как способ улучшения обобщающей способности нейросети в задаче детекции объектов. Труды ИСП РАН, том 35, вып. 6, 2023 г., стр. 247–264. DOI: 10.15514/ISPRAS-2023-35(6)-16.

Благодарности: Авторы выражают благодарность ОАО РЖД.

Style Transfer as a Way to Improve the Generalization Ability of a Neural Network in an Object Detection Task

^{1,2} D.K. Karachev, ORCID: 0000-0002-1008-2535 <denis.karachev@ocrv.ru>

¹ S.E. Shtekhin, ORCID: 0000-0003-2866-4864 <sergei.shtekhin@ocrv.ru>

² V.S. Tarasyan, ORCID: 0000-0002-5288-5221 <vladimir@usurt.ru>

¹ I.U. Smolin, ORCID: 0000-0002-7272-1232 <ilya.smolin@ocrv.ru>

¹ M.V. Isakov, ORCID: 0000-0002-8677-8862 <maksim.isakov@ocrv.ru>

¹ Industry Center for Information Systems' Development and Deployment,

2/1, Mejdunarodnaya, Sochi, 354340, Russia.

² Ural State University of Railway Transport,

66, Kolmogorova, Ekaterinburg, 620034, Russia.

Abstract. This paper proposes the implementation of a neural network training approach for object detection, using augmentation - style transfer. This method improves the generalization ability of the neural network to determine the location of objects in the image by improving the interaction with low-level features such as textures, different colors and small changes in shapes. The effectiveness of the method is experimentally proved and the numerical values of the object detection metrics are demonstrated on several datasets with different classes. The application of augmentation is proposed using an unused before neural network architecture capable of carrying an arbitrary number of styles. The peculiarity of the approach is also that the weights of the neural network for styling are frozen and it is added to the graph of the detection network, which allows augmentation speed.

Keywords: neural networks; computer vision; style transfer; machine learning; object detection.

For citation: Karachev D.K., Shtekhin S.E., Tarasyan V.S., Smolin I.U., Isakov M.V. Style transfer as a way to improve the generalization ability of a neural network in an object detection task. *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 6, 2023. pp. 247-264 (in Russian). DOI: 10.15514/ISPRAS-2023-35(6)-16.

Acknowledgements. The authors are grateful to the Russian Railway Company.

1. Введение

С 2012 года конволюционные нейронные сети начали широко применяться в области компьютерного зрения. По качеству предсказаний они превосходят классические алгоритмы компьютерного зрения. По скорости работы и объему потребляемой памяти выигрывают у полносвязных нейронных сетей. Однако в большинстве случаев в процессе обучения нейронной сети форма объектов не запоминается, и она предпочитает более лёгкий путь для достижения цели – переобучаться на текстуры [1].

Для решения этой проблемы, в задаче классификации изображений, были предложены различные методы переноса стиля [2]. Данные методы представлены в разделе 2.

Целью данной работы является применение метода переноса стиля изображений для повышения точности в задаче детектирования объектов в реальных условиях. Предлагаемый метод является универсальным для применения в различных областях, где решается задача детекции объектов. Для детектирования объектов использовалась архитектура нейронной сети Yolo v5 [3], в основе которой лежит конволюционная сеть. Метод был проверен на задаче детектирования инструментов на наборе данных с ограниченным доступом. При обучении нейронной сети выяснилось, что модель ошибается в распознавании некоторых инструментов, переобучаясь на похожие текстуры и цвета инструментов. Также метод был проверен на задаче детектирования кистей рук в перчатках разного цвета. Обучение такой модели проводилось на открытых наборах данных (рассмотрен в разделе 3) [4-7].

Разработка и применение аугментации переноса стиля для нейросети детектирования объектов позволили исключить переобучение нейросети на объекты с похожими текстурами

и разной формой. Применение аугментации предлагается реализовать с помощью ранее неиспользуемой архитектуры нейросети, способной переносить произвольное количество стилей. Особенностью подхода также является то, что веса нейросети для стилизации замораживаются, и нейронная сеть добавляется в граф детекционной сети, что позволяет ускорить скорость аугментации. Подробное описание метода представлено в разделе 3. Тестирование метода было проведено на данных, которые рассмотрены в разделе 4.

Применение аугментации переноса стиля позволило улучшить метрики детектирования объектов в среднем на 7%. Результаты экспериментов изложены в разделе 5. Обзор существующих методов представлен в следующем разделе.

2. Обзор существующих работ

Задача переноса стиля берет свое начало в рендеринге [8] и тесно связана с синтезом и передачей текстуры [9-11]. Некоторые ранние подходы включают сопоставление гистограмм по откликам линейных фильтров [12] и непараметрическую выборку [13, 14]. Эти методы обычно полагаются на низкоуровневую статистику и не способны уловить семантические структуры. Авторы работы [15] впервые продемонстрировали впечатляющие результаты переноса стиля путем сопоставления статистики признаков в конволюционных слоях глубокой нейронной сети. Недавно было предложено несколько усовершенствований. В работе [16] авторы представили схему, основанную на случайных марковских полях в пространстве признаков для обеспечения соблюдения локальных закономерностей. Также в [17] предложили способы управления изменением цвета, пространственным расположением и масштабом передачи стиля. В работе [18] улучшили качество передачи стиля видео, наложив временные ограничения.

Подход авторов [15] основан на медленном процессе оптимизации, который итеративно обновляет изображение, чтобы минимизировать функцию потерь для стиля и контента. Процесс переноса стиля данным методом занимает продолжительное время даже на современных графических процессорах. Решение данной проблемы заключается в замене процесса оптимизации на нейронную сеть, которая обучена оптимизировать ту же самую функцию [19-21]. Эти методы передачи информации стиля примерно на три порядка быстрее, чем альтернатива, основанная на оптимизации изображения, что открывает возможности для применения в реальном времени. В работе [22] предлагается алгоритм для улучшения гранулярности. Авторы в [23] предложили способы улучшения качества и разнообразия генерируемых образцов. Однако, вышеуказанные методы передачи стиля ограничены в том смысле, что каждая сеть привязана к фиксированному стилю. Для решения этой проблемы авторы работы [24] представили одну сеть, которая способна кодировать 32 стиля и их интерполяции. В работе [25] предложили прямолинейную архитектуру, которая может синтезировать до 300 текстур и перенос 16 стилей. Тем не менее, два вышеуказанных метода не могут адаптироваться к произвольным стилям, которые не наблюдаются во время обучения.

Авторы работы [26] представили метод передачи стиля, который может передавать произвольные стили благодаря специальному слою подмены. Учитывая активации признаков изображений контента и стиля, слой изменения стилей корректирует признаки контента на наиболее близкие по значению признаки стиля в режиме скользящего окна. Тем не менее, их слой замены стиля создает новое узкое место в вычислениях: более 95% вычислений тратится на корректировку стиля для входных изображений размера 512×512 .

Работа [15] потенциально может быть использована для обучения на изображениях. Из плюсов данного подхода можно отметить то, что он позволяет регулировать силу применения стиля за счёт изменения количества итераций. Однако это не позволяет быстро изменять стиль изображения в связи с идеей самого подхода. Он итеративно изменяет тензор и на каждой итерации пропускает картинки со стилем и контентом через нейросеть, что требует

значительного количества времени на обработку даже одного изображения. Поэтому данный вариант разумно использовать только когда имеется значительное количество памяти под набор данных, и есть возможность его сгенерировать заранее. Время, затраченное на стилизацию, будет прямо пропорционально зависеть от размера набора данных.

Стоит подробнее рассказать про работу [19], в которой автор предложил использовать идею из предыдущего исследования [15]. Автор предлагает новую функцию потерь под названием «потеря восприятия» (perceptual loss), с помощью которой решает задачи переноса стиля и повышение разрешения. Этот подход позволяет обучать отдельную нейросеть под каждый стиль. Следовательно, есть возможность на выходе получить небольшую нейросеть, которая имеет мало параметров и требует немного времени на исполнение. Недостаток подхода в том, что для того, чтобы иметь возможность применения большого количества стилей, необходимо большое количество нейросетей, что ограничивает вычислительные возможности и увеличивает время обучения сети. Также нужно тщательно выбирать стили, так как изначально нет информации о том, какой стиль приносит наибольшее влияние.

В работе [27] автор предлагает новый универсальный подход, который заключается в том, чтобы один раз обучить нейросеть, которая затем позволит перенести стиль из любого изображения. Аналогично авторам вышеупомянутых работ, предлагается использовать предобученную нейросеть типа VGG [28] для извлечения карт признаков изображений стиля и контента. Особенностью работы является блок под названием AdaIn (Adaptive Instance Normalization), благодаря которому происходит смешение стиля и контента в один набор карт признаков, который потом используется декодером, для получения итогового результата. Данный подход можно считать лучшим по сравнению с предыдущими, т.к. имеется возможность применять стилизацию в реальном времени для любого количества стилей без дополнительного дообучения нейросетей. Схема применения стилизации представлена на рис. 1.

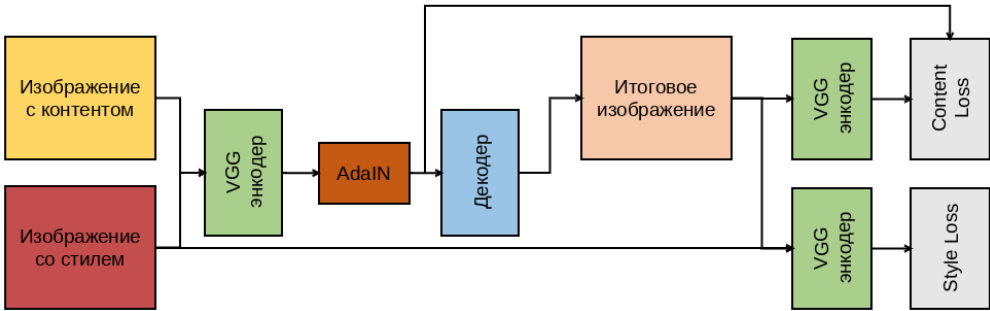


Рис. 1. Схема AdaIn подхода стилизации
Fig. 1. AdaIn's diagram of the stylization approach

Другой проблемой при передаче стиля является выбор функции потерь. Подход [15] обеспечивает соответствие стилей путем сопоставления статистики второго порядка между активациями признаков, отражаемой матрицей Грамма. Были предложены и другие эффективные функции потерь, такие как «adversarial loss» [21], «histogram loss» [29], «CORAL loss» [30] и расстояние между средним и дисперсией канала [31]. Стоит отметить, что все вышеперечисленные функции потерь направлены на согласование некоторых статистических характеристик между образцом и синтезированным изображением.

Существует несколько альтернативных подходов генерации изображений, включая вариационные автокодировщики [32], авторегрессионные модели [33] и генеративно-сопоставительные сети (GAN) [34]. Примечательно, что GAN достигли наиболее

впечатляющего визуального качества. GAN также применялись для передачи стиля [14] и генерации изображений для кросс доменных преобразований [35-37].

В работе [2] преобразования стиля используются для аугментации при обучении нейронной сети для задачи классификации. Авторы используют несколько разновидностей стиля и производят их сравнения. В результате они получают значительное преимущество при использовании совместно с классическими аугментациями.

Работа [38] подчёркивает проблематику детекционных нейросетей с объектами, которые не находились в домене обучающей выборки. Авторы используют технику переноса стиля для обучения нейросети, которая должна детектировать людей на картинах различных художников. В работе применяется классический AdaIn [27] подход для генерации искусственного набора данных styleCOCO, из части открытого набора данных Microsoft's COCO [41], на котором происходит дообучение предобученной модели Faster-RCNN и достигается лучший результат на тестовом наборе данных People-Art [39]. Данная работа подтверждает актуальность использования подходов изменения стиля изображений для обучения нейросетей. Стоит отметить, что авторы не ставили целью улучшить детекцию на реальных объектах и ограничились только исследованиями в области художественных картин.

Вышеприведенные работы использовали стилизацию в основном для задач классификации за исключением [38], в которой создается styleCOCO набор данных для улучшения работы детектора на картинах художников. Стоит отметить, что как модель стилизации изображений для создания обучающего датасета, так и целевые доменные области работы детектора отличаются в сравнении с текущей работой. В статье [38] использовалась стилизация, чтобы реальные люди на фотографиях, в оригинальном датасете COCO, стали похожи на людей на картинах, таким образом получается, что стилизованный датасет styleCOCO - это своеобразная вариация датасета People Art. За счет применения стилизации эти датасеты становятся визуально похожи. В данной работе, стилизация применяется для обучения нейросети на объектах, которые не похожи на тот набор объектов, которые нейросеть должна определять. Обученная модель должна детектировать реальные инструменты и перчатки, которые не похожи визуально на стилизованные.

В данной работе представлены результаты исследований применительно к задаче детекции для нескольких наборов данных, собранных в реальных условиях, которые, как специализированные, так и находящиеся в открытом доступе. Подробное описание предложенного метода, представлено далее.

3. Предлагаемый метод

В предыдущем разделе работы были рассмотрены существующие методы борьбы с переобучением и повышения устойчивости работы нейросетей со своими преимуществами и недостатками. Здесь предлагается использовать перенос стиля не только как метод для борьбы с переобучением, но еще и как метод для улучшения обобщающей способности нейросети в случае, когда объекты могут отличаться, например, по цветам. В качестве наглядного примера можно привести детекцию кистей рук. Вариант, когда в обучающей выборке все изображения, на которых руки без перчаток, но тестирование происходит на картинках, на которых перчатки присутствуют. Изменение стиля позволит получить вариативность цветовой окраски объектов на изображении и обеспечит незначительные изменения в формах объектов, что искусственно расширит обучающий набор данных и позволит в дальнейшем лучше детектировать кисти рук в перчатках.

На текущем этапе работы в качестве детектора используется нейросеть типа YOLOv5 [3]. Она была выбрана, так как на открытых данных имеет хорошие метрики скорости и точности, однако в наших тестовых данных имеются существенные недочеты и ошибки в предсказаниях. Основными проблемами является то, что модель путает классы между собой,

а также не детектирует один и тот же класс на изображении, при изменении цвета детектируемого объекта. Для устранения этих ошибок было предложено использовать стилизованную аугментацию.

В настоящее время существует несколько алгоритмов переноса стиля, каждый из них имеет свои плюсы и минусы. В зависимости от метода стилизации изменяется итоговый алгоритм обучения модели детекции.

Для применения стилизации был выбран подход, аналогичный работе [28], т.к. он имеет ряд преимуществ перед другими подходами. Однако была сделана небольшая модификация. Энкодер нейросети VGG дополнительно связывается с декодером с помощью дополнительных конкатенаций между слоями одного уровня по типу реализации U-Net [39] сети, но с дополнительными AdaIn блоками между ними. Упрощенная схема модифицированной сети представлена на рис. 2.

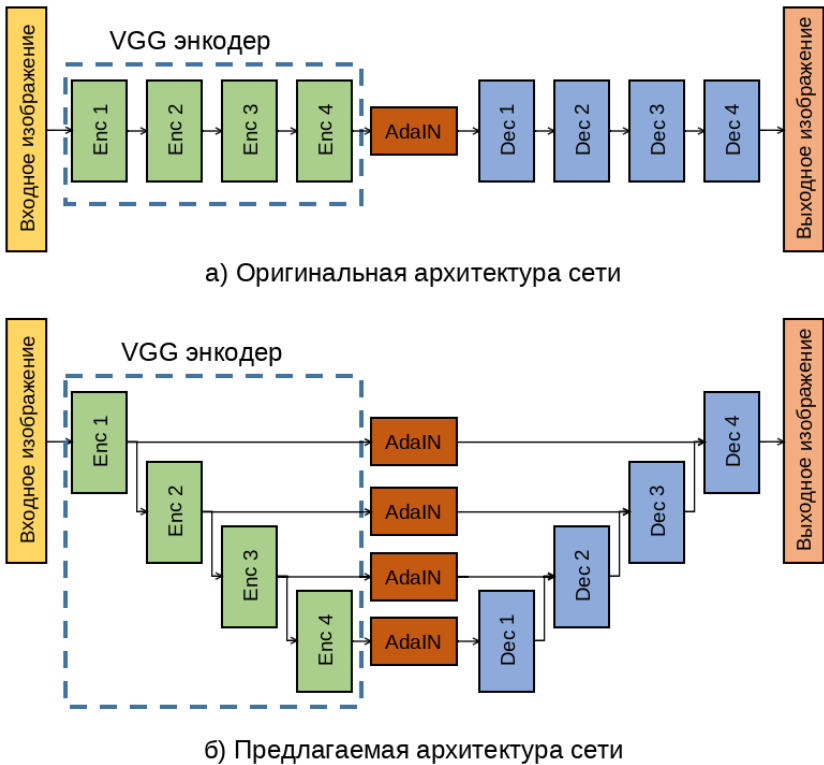


Рис. 2. Оригинальная и предлагаемая архитектура нейросети для стилизации изображений
Fig. 2. Original and modified neural network architecture for image styling

Обычно процесс аугментации изображения происходит с использованием нескольких потоков во время подготовки пакета данных, перед его подачей в нейросеть. Однако данный подход не является подходящим для предлагаемой аугментации. Стилизованные изображения для обучения нейросети можно получить несколькими способами:

- сгенерировать набор данных заранее;
- изменять изображения в процессе формирования пакета данных.

Генерация набора данных заранее является плохой практикой. Во-первых, для корректной работы аугментации, необходимо для каждой обучающей эпохи сгенерировать набор изображений, который по количеству изображений будет равен исходной тренировочной

выборке. Это значит, что память, занимаемая изображениями, на жестком диске увеличится в N раз, где N – количество обучающих эпох. Также необходимо изменять конвейер (pipeline) обучения нейросети, для того чтобы правильно подгружать необходимые картинки для каждой эпохи. Поэтому было решено, что данный подход не подходит для текущей реализации.

Классическим подходом для реализации аугментаций является изменение изображений в процессе подготовки пакета данных. Во время процесса обучения нейросети происходит сбор изображений в режиме многопоточности, что позволяет быстро собирать наборы картинок и подавать их в нейросеть. Базовые аугментации, осуществляющие преобразования поворота и масштабирования, применяются во время этого этапа предобработки изображений, требуют мало вычислительных ресурсов, быстро работают и не требуют много дополнительной памяти. Если использовать стилизацию аналогичным способом, то необходимо создавать экземпляр нейросети для переноса стиля в каждом потоке, что займет дополнительную память. Также произойдет замедление работы за счет того, что операции по сборке пакета данных происходят с использованием центрального процессора. Использование графических ускорителей для аугментации при сборке пакета данных изображений нецелесообразно. На графических картах происходит обучение детекционной нейросети и требуется вся видеопамять для нейросети и пакета данных. Чем больше потоков для сбора пакета данных используется, тем больше видеопамати необходимо для аугментации.

Исходя из вышесказанного, можно сделать вывод, что использовать нейронную сеть для стилизации в процессе сборки пакета данных нецелесообразно. В данной работе предлагается объединить детекционную нейросеть и нейросеть для стилизации в один граф, заморозив при этом веса нейросети для стилизации. Стилизация будет применяться с определенной вероятностью для всего пакета данных во время обучения нейросети для детекции. Таким образом, решается проблема распараллеливания на нескольких графических устройствах, т.к. количество экземпляров нейросетей для стилизации равно количеству нейросетей для детекции, и они считаются за один экземпляр нейросети. Вследствие этого происходит экономия видеопамати. Отсутствует привязка к количеству потоков при сборе пакета данных. Также скорость работы аугментации значительно увеличивается за счёт графических ускорителей. Схема реализованного подхода аугментации данных представлена на рис. 3. Наборы данных, на которых тестировался реализованный подход, описаны в следующем разделе.

4. Наборы данных

Основной набор данных для задачи детектирования объектов представляет собой набор из 200 аннотированных видеофайлов с изображениями 36 специализированными инструментами. Средняя продолжительность видео от 8 до 10 минут, частота кадров – 25, разрешение видео 1280x720, 1080x1920.

В данной работе, для проверки гипотезы об улучшении качества обучения детекционной сети с помощью аугментации из основного набора данных были выбраны три контрольные подвыборки (табл. 1), содержащие изображения с:

- двумя инструментами (визуально схожими);
- шестью инструментами (неуверенно распознаваемыми);
- шестью инструментами, из которых три неуверенно и три хорошо распознаваемые.

Первая подвыборка содержит два инструмента и является базовой для подбора параметров аугментации. Вторая и третья подвыборка содержат по шесть инструментов и являются контрольными для проверки качества ошибочно детектируемых инструментов.



Рис. 3. Схема реализованного подхода применения аугментации
Fig. 3. Scheme of the implemented augmentation approach

Табл. 1. Названия и соответствующие им индексы инструментов
Table 1. Names and corresponding tool indices

Набор инструментов	Индекс	Название рус.	Название англ.
1	0	ключ торцевой	socket wrench
	1	метла	broom
2	0	ключ торцевой	socket wrench
	1	метла	broom
	2	молоток костыльный	spike maul
	3	кувалда	sledge hammer
	4	лопата	shovel
	5	трамбовка ручная	wooden rammer
3	0	ключ торцевой	socket wrench
	1	метла	broom
	2	домкрат гидравлический	hydraulic jack
	3	шаблон путевой	track template
	4	станок рельсошлифовальный	rail grinder small
	5	кувалда	sledge hammer

Обусловлено это тем, что из-за большого количества данных обучение на всех инструментах занимает большое количество времени, что делает невозможным подбор всех параметров для обучения за разумное время. Выбор инструментов происходил таким образом, чтобы

проверить улучшение детекции ошибочно классифицируемых инструментов (например, лопата и торцевой ключ), при этом не ухудшив качество хорошо детектируемых инструментов.

Баланс классов и информация об ограничивающих прямоугольниках представлены на рис. 4 и рис. 5 соответственно.

На каждый класс приходится от 8 до 11 тыс. объектов, преимущественно они сосредоточены в центре изображения, но есть и в других произвольных местах. Размер ограничивающих прямоугольников в основном составляет 10-40% от размеров всего изображения, но также есть экземпляры, занимающие большую часть изображения.



Рис. 4. Баланс классов для трёх наборов данных
Fig. 4. Class balance for three datasets

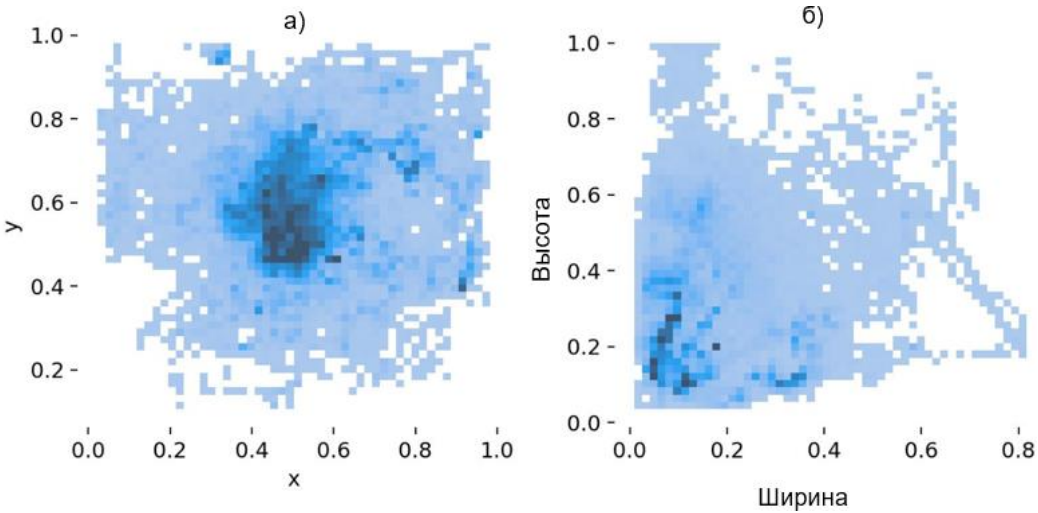


Рис. 5. Координаты центра ограничивающих прямоугольников (а) и их размеры относительно размеров изображения (б)
Fig. 5. The coordinates of the center of the bounding boxes (a) and their dimensions relative to the dimensions of the image (b)

Дополнительно, обучение моделей для детектирования кистей рук человека проводилось на наборах данных, представленных в свободном доступе:

- Human-Parts [4], представляет собой набор данных с тремя классами: человек, лицо человека, кисть руки человека.

- HaGRID (HAnd Gesture Recognition Image Dataset) [5], набор данных для распознавания жестов, снятый на веб-камеру и камеру телефона.
- COCO-hands [6], 26 тыс. изображений, содержащих кисть руки человека из набора данных Microsoft's COCO [41].
- TV-hands, набор изображений извлеченных из набора данных ActionThread [42], который содержит около 4500 видеороликов. Из каждого ролика было выбрано 2-3 кадра.
- YouTube-hands [7], нарезка кадров видео с платформы YouTube. Представлены различные действия: приготовление еды, танцы, спортзал, игра на музыкальных инструментах, работа с инструментами и т.д. Для объективной оценки качества детекции, видео, на которых руки были в перчатках исключены.

Обучение нейросети для переноса стиля происходило на наборе данных MS-COCO [41], который использовался как изображения контента, изображения для стилизации были взяты из набора данных WikiArt [43]. Параметры экспериментов, которые были проведены на описанных наборах данных и полученные результаты экспериментов, подробно представлены в следующем разделе.

5. Эксперименты и сравнительный анализ полученных результатов

Обучение производилось на одной видеокарте Nvidia A100 40 GB. Следующие гиперпараметры были выбраны для обучения: размер пакета данных (batch_size) 32, количество потоков (num_workers) 16, коэффициент скорости обучения (learning_rate) 0.0001, коэффициент стиля (style_weight) 5.0, коэффициент контента (content_weight) 1.0. В качестве оценки точности была выбрана метрика mean Average Precision (mAP) [44] с использованием порогового значения Intersection over Union 0.5 (mAP 0.5) и среднее значение mAP 0.5..0.95 по разным порогам Intersection over Union, от 0.5 до 0.95 [45].

Результат применения нейросети для стилизации изображений представлен на рис. 6.

Обучение детекционной сети на инструментах происходит на наборе данных с ограниченным доступом, подробная информация о котором представлена выше. Всего для обучения доступны 36 инструментов. Прежде чем проверять гипотезу с использованием аугментации, модель детекции была обучена на всех инструментах.

Детекционная нейросеть обучалась со следующими параметрами:

1. epoch – 96;
2. batch_size – 12;
3. img_size – 1024x1024;
4. final learning rate – 0.001;
5. optimizer – SGD;
6. средние аугментации семейства Yolo.

Из табл. 2, где представлены метрика качества mAP для сравнения подходов, видно, что использование стилизации как аугментации улучшает метрику детекции от 0.004 до 0.05 mAP, что свидетельствует о значительном положительном эффекте использования стилизации.



Рис. 6. Пример оригинальных и стилизованных изображений из нашего набора данных с измененным инструментом

Fig. 6. Example of original and stylized images from our dataset with a modified tool

Табл. 2. Результаты обучения детекционной нейросети на инструментах с использованием аугментации для изменения стиля и без нее

Table 2. Training results of a detection neural network on instruments with/without style augmentation

	Обычные аугментации		Обычные аугментации + стилизация	
	mAP 0.5	mAP 0.5..0.95	mAP 0.5	mAP 0.5..0.95
1 набор (2 инструмента)	0.878	0.486	0.925	0.547
2 набор (6 инструментов)	0.838	0.568	0.876	0.588
3 набор (6 инструментов)	0.851	0.576	0.855	0.577

Также, после проведения экспериментов было выявлено, что стилизация позволяет избежать ошибок предсказаний и значительно повышает качество детекции. При этом качество

предсказания на других инструментах не ухудшилось. На рис. 7а), 7в) представлены примеры работы модели детектирования инструментов, обученной с применением алгоритма стилизации, на рис. 7б), 7г) – результаты работы модели детектирования инструментов, обученной без использования алгоритма стилизации. Как видно из рисунков, модель, обученная с применением аугментации, детектирует инструменты, которые ранее не обнаруживались.



Рис. 7. Пример детекции инструментов
Fig. 7. Tool detection example

Аугментация с изменением стиля также была протестирована в задаче детекции кистей рук, так как было замечено, что модель плохо отрабатывает в случаях, когда человек в перчатках. Причиной такой работы нейросети является то, что открытые данные, на которых она обучалась, не содержат изображения с кистями рук в перчатках. Изображения в перчатках изменяют не только цвет рук, но и объем. Использование стилизации в данном случае является подходящим решением, так как она изменяет цвет текстуры, а также производит небольшую деформацию, в зависимости от применяемого стиля. Пример изменения стиля представлен на рис. 8.

Детекционная модель обучалась на нескольких наборах данных [4-7], которые находится в открытом доступе. Параметры обучения сети аналогичны эксперименту с инструментами, описанному выше.

Результаты обучения нейросети на кисти рук приведены в табл. 3.

По табл. 3 видны значительные улучшения метрики качества детекции от 0.01 до 0.19 mAP, что подтверждает эффективность использования аугментации по изменению стиля. На рис. 9а), 9в) представлены примеры работы модели детектирования кистей рук, обученной с применением алгоритма стилизации, на рис. 9б), 9г) – результаты работы модели детектирования кистей рук, обученной без использования алгоритма стилизации. Как видно из рисунков, модель, обученная с применением аугментации, детектирует кисти рук, которые ранее не обнаруживались.

Визуальный пример сравнительного предсказания моделей представлен на рис. 9.

Табл. 3. Результаты обучения детекционной нейросети на кисти рук с использованием аугментации для изменения стиля и без нее

Table 3. Training results of a detection neural network on hands with/without style augmentation

Набор данных	Обычные аугментации		Обычные аугментации + стилизация	
	mAP 0.5	mAP 0.5..0.95	mAP 0.5	mAP 0.5..0.95
Human-Parts [4]	0.81	0.50	0.82	0.50
HaGRID [5]	0.53	0.27	0.54	0.28
COCO-hand [6]	0.77	0.37	0.83	0.38
TV-hand [40]	0.57	0.24	0.67	0.26
YouTube-hand [7]	0.70	0.42	0.89	0.57
Все наборы данных	0.84	0.49	0.9	0.51

6. Заключение

В данной работе предложен подход по улучшению обобщающей способности нейросети для задачи детекции объектов (Object Detection) с помощью метода переноса стиля изображений. Данный метод улучшает обобщающую способность нейросети для определения местоположения объектов на изображении путём уменьшения переобучения нейросети на низкоуровневые признаки, такие как текстуры, цвет и небольшие отклонения форм. Предложена ранее не использованная модификация архитектуры нейросети для стилизации изображений, которая является существенным вкладом авторов в данную проблематику. Разработана программная реализация предлагаемого метода для использования в качестве аугментации, способной переносить произвольное количество стилей на изображения.

Экспериментально подтверждено улучшение обобщающей способности нейросети на нескольких наборах данных из области детекции частей тела (Human-Parts, HaGRID, COCO-hand и др.) Полученные результаты демонстрируют универсальность предлагаемого метода и устойчивость работы в задачах различного прикладного назначения. На наборе данных с ограниченным доступом для детектирования железнодорожных инструментов было достигнуто улучшение метрики mAP@0.5 на 4.5%. В среднем было достигнуто улучшение метрики mAP@0.5 на 7% как для наборов данных с ограниченным доступом к инструментам, так и для открытых наборов данных для распознавания кистей рук.



Рис. 8. Пример изображения обычной руки и стилизованной
Fig. 8. Example with a normal hand and a stylized hand

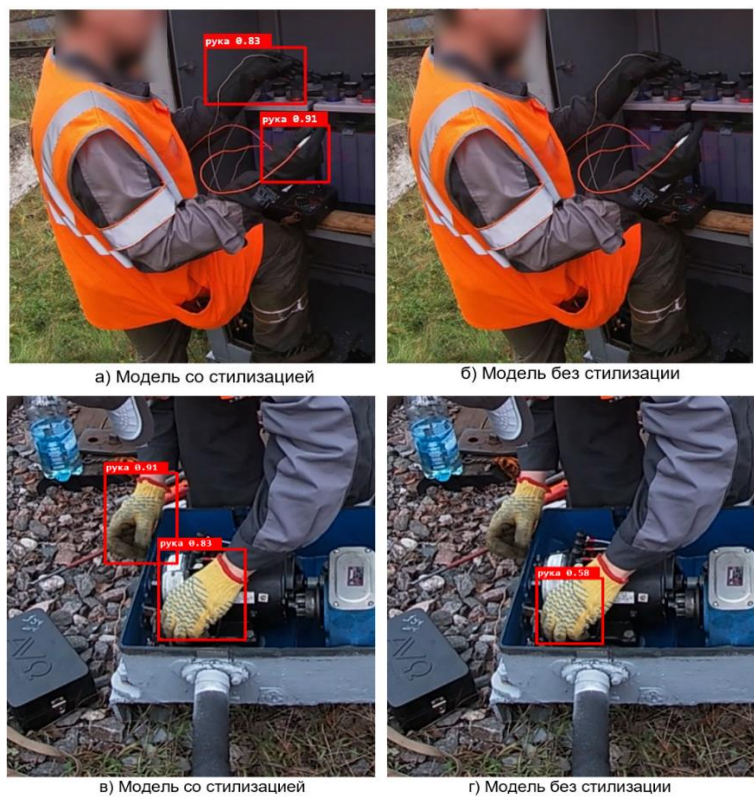


Рис. 9. Пример детекции кистей рук
Fig. 9. Hand detection example

Список литературы / References

- [1]. Гримов Р. Нейронные сети предпочитают текстуры и как с этим бороться. Хабр. <https://habr.com/ru/company/ods/blog/453788/>.
- [2]. Xu Zheng, Tejo Chalasani, Koustav Ghosal, Sebastian Lutz, and Aljosa Smolic. STaDA: Style Transfer as Data Augmentation. In Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Vol. 4 VISAPP: VISAPP, 107-114, 2019, Prague, Czech Republic, <https://www.scitepress.org/Link.aspx?doi=10.5220/0007353401070114>.
- [3]. Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., ... Rai, P. ultralytics/yolov5: v3.1 – Bug Fixes and Performance Improvements (v3.1) [Computer software]. Zenodo, 2020. <https://doi.org/10.5281/zenodo.4154370>.
- [4]. Xiaojie Li., Lu Yang, Qing Song, Fuqiang Zhou. Detector-in-Detector: Multi-Level Analysis for Human Parts, 2019. <https://arxiv.org/abs/1902.07017>.
- [5]. Kapitanov A., Makhlyarchuk A., Kvanchiani K. HaGRID. Hand Gesture Recognition Image Dataset. SberDevices, Russia, 2022, arXiv:2206.08219v1.
- [6]. Narasimhaswamy S., Wei Z., Wang Y., Zhang J., Hoai M., Contextual Attention for Hand Detection in the Wild. IEEE/CVF International Conference on Computer Vision (ICCV), 2019. doi: 10.1109/ICCV.2019.00966.
- [7]. Huang, M., Narasimhaswamy, S., Vazir, S., Ling, H., Hoai, M. Forward Propagation, Backward Regression and Pose Association for Hand Tracking in the Wild. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 6396-6406, doi: 10.1109/CVPR52688.2022.00630.
- [8]. Kyprianidis, J. E., Collomosse, J., Wang, T., & Isenberg, T. State of the ‘Art’: A Taxonomy of Artistic Stylization Techniques for Images and Video. IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 5, pp. 866-885, May 2013. doi: 10.1109/TVCG.2012.160.
- [9]. Efros A. A. and Leung T. K. Texture synthesis by non-parametric sampling. Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 1999, pp. 1033-1038 vol.2, doi: 10.1109/ICCV.1999.790383.
- [10]. Lara Raad, and Bruno Galerne, Efros and Freeman Image Quilting Algorithm for Texture Synthesis, Image Processing On Line, 7 (2017), pp. 1–22. doi:10.5201/ipol.2017.171.
- [11]. Elad M. and Milanfar P. Style-transfer via texture-synthesis. arXiv. doi:10.48550/arXiv.1609.03057.
- [12]. Heeger D. J., Bergen J. R. Pyramid-Based Texture Analysis/Synthesis. Proceedings of the International Conference on Image Processing (SIGGRAPH), 1995, vol. 3, pp. 648-651, <https://api.semanticscholar.org/CorpusID:47266338>.
- [13]. Bousmalis K., Silberman N., Dohan D., Erhan D., Krishnan D. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. arXiv:1612.05424v1, 2016.
- [14]. Collobert R., Kavukcuoglu K., Farabet C. Torch7: A Matlab-like Environment for Machine Learning. In: BigLearn, NIPS Workshop, 2011.
- [15]. Gatys L. A., Ecker A. S., Bethge M. Image style transfer using convolutional neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2414-2423, doi: 10.1109/CVPR.2016.265.
- [16]. Li C., Wand M. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2479-2486, <https://doi.org/10.1109/CVPR.2016.272>.
- [17]. Gatys L. A., Ecker A. S., Bethge M., Hertzmann A., Shechtman E. Controlling Perceptual Factors in Neural Style Transfer. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3730-3738, doi: 10.1109/CVPR.2017.397.
- [18]. Ruder M., Dosovitskiy A., Brox T. Artistic Style Transfer for Videos. In: Rosenhahn B., Andres B., (eds). Pattern Recognition. GCPR 2016. Lecture Notes in Computer Science(), vol 9796. Springer, Cham. https://doi.org/10.1007/978-3-319-45886-1_3.
- [19]. Johnson J., Alahi A., Fei-Fei L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9906. Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6_43.

- [20]. Ulyanov D., Lebedev V., Vedaldi A., Lempitsky V. Texture networks: Feed-forward Synthesis of Textures and Stylized Images. Proceedings of the 33-rd International Conference on Machine Learning (ICML), New York, NY, USA, 2016. JMLR: W&CP vol. 48, pp. 2027-2041, 2016.
- [21]. Li C., Wand M. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9907. Springer, Cham. https://doi.org/10.1007/978-3-319-46487-9_43.
- [22]. Wang X., Oxholm G., Zhang D., Wang Y.-F. Multimodal Transfer: A Hierarchical Deep Convolutional Neural Network for Fast Artistic Style Transfer. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 7178-7186, doi: 10.1109/CVPR.2017.759.
- [23]. Ulyanov D., Vedaldi A., Lempitsky V. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 4105-4113, doi: 10.1109/CVPR.2017.437.
- [24]. Dumoulin V., Shlens J., Kudlur M. A Learned Representation for Artistic Style., In Proceedings of the International Conference on Learning Representations (ICLR), 2017, <https://openreview.net/forum?id=BJO-BuT1g>, arXiv preprint arXiv:1610.07629, 2016.
- [25]. Li Y., Fang C., Yang J., Wang Z., Lu X., Yang M.-H. Diversified Texture Synthesis with Feed-Forward Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 266-274, doi: 10.1109/CVPR.2017.36.
- [26]. Chen T. Q., Schmidt M. Fast Patch-based Style Transfer of Arbitrary Style. arXiv preprint arXiv:1612.04337, 2016.
- [27]. Huang, X., Belongie, S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization, 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 1510-1519, doi: 10.1109/ICCV.2017.167.
- [28]. Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014. arXiv. doi: 10.48550/arXiv.1409.1556
- [29]. Wilmot P., Risser E., Barnes C. Stable and Controllable Neural Texture Synthesis and Style Transfer Using Histogram Losses. arXiv. preprint arXiv:1701.08893, 2017.
- [30]. Peng X., Saenko K. Synthetic to Real Adaptation with Deep Generative Correlation Alignment Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 2018, pp. 1982-1991, doi: 10.1109/WACV.2018.00219.
- [31]. Li Y., Wang N., Liu J., Hou X. Demystifying Neural Style Transfer. arXiv. 2017. doi: 10.48550/arXiv.1701.01036.
- [32]. Kingma D. P., Welling M. Auto-encoding variational bayes. In ICLR, 2014, doi: 10.48550/arXiv.1312.6114.
- [33]. Oord A. v. d., Kalchbrenner N., Kavukcuoglu K. Pixel recurrent neural networks. In ICML, 2016, doi: 10.48550/arXiv.1601.06759.
- [34]. Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative adversarial nets. NIPS 2014, 2672; <http://papers.nips.cc/paper/5423-generative-adversarial-nets>.
- [35]. Taigman Y., Polyak A., Wolf L. Unsupervised cross- domain image generation. In ICLR, 2017, doi: 10.48550/arXiv.1611.02200.
- [36]. Liu M.-Y., Tuzel O. Coupled generative adversarial networks. In NIPS, 2016, <https://proceedings.neurips.cc/paper/2016/file/502e4a16930e414107ee22b6198c578f-Paper.pdf>.
- [37]. Kim T., Cha M., Kim H., Lee J., Kim J. Learning to discover cross-domain relations with generative adversarial networks. arXiv, 2017, doi: 10.48550/arXiv.1703.05192.
- [38]. Kadish D., Risi S., Løvlie A. S. Improving object detection in art images using only style transfer //2021 International Joint Conference on Neural Networks (IJCNN). – IEEE, 2021. – C. 1-8.
- [39]. Cai H. et al. The cross-depiction problem: Computer vision algorithms for recognising objects in artwork and in photographs //arXiv preprint arXiv:1505.00110. – 2015.
- [40]. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv 2015. doi: 10.48550/arXiv.1505.04597.
- [41]. Lin T.-Y., Maire M., Belongie S., Bourdev L., Girshick R., Hays J., Perona P., Ramanan D., Zitnick C. L., Dollár P. (2014). Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B.,

Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol. 8693. Springer, Cham. doi: 10.1007/978-3-319-10602-1_48.

- [42]. Wang Y., Hoai, M. Improving Human Action Recognition by Non-action Classification. arXiv 2016. doi: 10.48550/arXiv.1604.06397.
- [43]. Tan W. R., Chan C. S., Aguirre H., Tanaka K. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. IEEE Transactions on Image Processing, 28(1), pp. 394–409, 2019. doi: 10.1109/TIP.2018.2866698.
- [44]. Beitzel S.M., Jensen E.C., Frieder O. MAP. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, 2009. doi: 10.1007/978-0-387-39940-9_492.
- [45]. Rezaatofghi H., Tsoi N., Gwak J., Sadeghian A., Reid I., Savarese S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. arXiv. 2019. doi: 10.48550/arXiv.1902.09630.

Информация об авторах / Information about authors

Денис Константинович КАРАЧЕВ – ведущий специалист по анализу данных. Сфера научных интересов: компьютерное зрение.

Denis Konstantinovich KARACHEV – Senior Data Scientist. Research interests: computer vision.

Сергей Евгеньевич ШТЕХИН – Руководитель группы компьютерного зрения лаборатории ИИ и нейронных сетей. Сфера научных интересов: компьютерное зрение.

Sergey Evgenievich SHTEKHIN – Team Lead Data Scientist. Research interests: computer vision.

Владимир Сергеевич ТАРАСЯН – кандидат физ.-мат. наук, доцент Уральского государственного университета путей сообщения, заведующий кафедрой “Мехатроника”. Сфера научных интересов: компьютерное зрение, интеллектуальный анализ данных, интеллектуальные системы управления.

Vladimir Sergeevich TARASYAN – Cand. Sci. (Phys.-Math.), Associate Professor at Ural state University of railway transport, Head of Mechatronic Department. Research interests: computer vision, intelligence data analysis, intelligence control systems.

Илья Юрьевич СМОЛИН – старший специалист по анализу данных. Сфера научных интересов: компьютерное зрение.

Ilya Urevich SMOLIN – Middle Data Scientist. Research interests: computer vision.

Максим Владимирович ИСАКОВ – специалист по анализу данных. Сфера научных интересов: компьютерное зрение.

Maksim Vladimirovich ISAKOV – Junior Data Scientist. Research interests: computer vision.

