МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
базового высшего образования
по направлению подготовки
09.03.02 Информационные системы и технологии,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Подготовка и разметка данных

Направление подготовки: 09.03.02 Информационные системы и технологии

Направленность (профиль): Технологии искусственного интеллекта в транспортных системах

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Дисциплина формирует фундаментальные и прикладные компетенции в области инженерии данных для задач машинного обучения на транспорте. В условиях перехода отрасли к интеллектуальным системам управления и автономному вождению критически востребованы специалисты, способные выстроить воспроизводимый пайплайн от сырых мультимодальных потоков до валидированных аннотированных датасетов. Студенты осваивают полный жизненный цикл подготовки данных – от анализа бизнес-требований и проектирования схем разметки до очистки телеметрии, работы с инструментами аннотирования и контроля качества. Практическое ядро курса опирается на импортозамещенный стек и открытые решения, что гарантирует готовность выпускников к работе в закрытых корпоративных контурах. Итогом обучения является самостоятельная реализация курсового проекта, результатом которого становится готовый к обучению моделей датасет с полным комплектом технической документации.

Целью освоения дисциплины является формирование у обучающихся системных знаний и практических умений в области сбора, инженерной подготовки, аннотирования и контроля качества мультимодальных данных для решения прикладных задач машинного обучения в транспортных системах.

Для достижения поставленной цели в рамках дисциплины решается комплекс задач, направленных на формирование у обучающихся способности: анализировать требования транспортных задач и проектировать оптимальные схемы разметки объектов, разрабатывать технические инструкции для аннотаторов, выполнять программную очистку и синхронизацию видеопотоков с логами бортовых систем, применять специализированное программное обеспечение для разметки данных, проводить статистический контроль качества аннотаций с расчетом метрик согласованности, а также версионировать датасеты и оформлять техническую документацию в соответствии с государственными стандартами.

2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

ПК-7 - Способен осуществлять сбор, подготовку, разметку и анализ данных для обучения моделей искусственного интеллекта.

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

Знать:

- жизненный цикл данных в проектах машинного обучения от сбора сырых массивов до эксплуатации обученной модели;
- типологию задач машинного обучения и их специфические требования к структуре и качеству данных;
- этические принципы и нормативные ограничения при работе с данными, включая методы обезличивания персональных данных;
- геометрические примитивы разметки в задачах компьютерного зрения – bounding box, polygon, polyline и семантическую сегментацию;
- схемы аннотирования неструктурированных текстовых данных и временных рядов телеметрии;
- принципы проектирования иерархии классов и таксономии объектов для транспортных задач;
- структуру и требования к техническому гайдлайну разметчика с разбором граничных случаев;
- форматы представления сырых мультимодальных данных и методы их первичной автоматической валидации;
- инженерные методы очистки, нормализации и синхронизации видеопотоков с логами бортовых систем;
- архитектуру и системные требования self-hosted инструментов разметки cvat и label studio;
- эргономику процесса аннотации и стратегии фиксации спорных объектов;
- метрики оценки качества разметки – intersection over union и коэффициент каппа коэна;
- статистические методы анализа распределения классов и выявления системных смещений в датасете;
- стратегии балансировки классов и стратификации выборок для обучения моделей;
- структурные особенности индустриальных форматов экспорта датасетов coco, yolo и pascal voc;
- принципы версионирования крупных массивов данных с использованием data version control;
- структуру паспорта датасета и описание ограничений его использования;
- требования государственных стандартов к оформлению пояснительных записок и технической документации.

Уметь:

- проектировать схему разметки данных и иерархию классов при помощи методов декомпозиции ML-задач при условии соответствия выбранной схемы целевой транспортной задаче;

- разрабатывать техническую инструкцию для разметчиков при помощи отечественного офисного пакета р7-офис при условии обязательного включения визуальных иллюстраций и разбора граничных случаев;

- выполнять предварительную обработку и синхронизацию сырых мультимодальных данных при помощи библиотек python и opencv при условии сохранения полной воспроизводимости всех трансформаций;

- выполнять аннотацию изображений и видеопоследовательностей при помощи инструментов cvat или label studio при условии строгого следования разработанному гайдлайну и соблюдения этических норм;

- проводить процедуру контроля качества разметки с расчетом метрик согласованности при помощи python-скриптов и библиотек статистического анализа при условии обязательного выявления системных смещений;

- выполнять балансировку классов и фильтрацию некачественных аннотаций при помощи методов статистического анализа при условии сохранения репрезентативности выборки по ключевым признакам;

- экспортировать готовый датасет в стандартные форматы и версионировать его при помощи git и dvc при условии обеспечения совместимости с современными фреймворками глубокого обучения;

- формировать паспорт датасета и итоговую пояснительную записку при помощи отечественного офисного пакета при условии полного описания архитектуры пайплайна и этических аспектов.

Владеть:

- навыками работы в специализированных средах аннотирования данных и инструментами версионирования датасетов;

- методами написания скриптов для автоматической валидации геометрии аннотаций и расчета метрик качества;

- приемами синхронизации видеопотоков с логами бортовых систем транспорта для комплексной разметки;

- способностью оформлять техническую документацию и паспорта датасетов в соответствии с требованиями ГОСТ.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 4 з.е. (144 академических часа(ов).

3.2. Объем дисциплины (модуля) в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Тип учебных занятий	Количество часов	
	Всего	Семестр №5
Контактная работа при проведении учебных занятий (всего):	64	64
В том числе:		
Занятия лекционного типа	32	32
Занятия семинарского типа	32	32

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 80 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

4. Содержание дисциплины (модуля).

4.1. Занятия лекционного типа.

№ п/п	Тематика лекционных занятий / краткое содержание
1	Введение в инженерию данных для машинного обучения Рассматриваемые вопросы: - жизненный цикл данных в проектах машинного обучения от сбора до эксплуатации модели; - типология задач машинного обучения и их специфические требования к структуре данных.
2	Этические и нормативные ограничения при работе с транспортными данными Рассматриваемые вопросы: - принципы конфиденциальности и методы обезличивания персональных данных; - предотвращение системных смещений и обеспечение репрезентативности выборок.

№ п/п	Тематика лекционных занятий / краткое содержание
3	Типология аннотаций в задачах компьютерного зрения Рассматриваемые вопросы: - геометрические примитивы разметки, включая bounding box, polygon и polyline; - семантическая и инстанс сегментация изображений и видеопоследовательностей.
4	Схемы разметки неструктурированных данных и временных рядов Рассматриваемые вопросы: - классификация документов и выделение именованных сущностей в текстах; - аннотирование событий и аномалий в логах телеметрии и временных рядах.
5	Проектирование таксономии и иерархии классов объектов Рассматриваемые вопросы: - принципы декомпозиции предметной области на логические классы; - правила вложенности, взаимного исключения и атрибутизации объектов.
6	Методология составления технического гайдлайна для разметчиков Рассматриваемые вопросы: - структура инструкции и требования к визуальным иллюстрациям; - разбор граничных случаев и критерии приемки аннотаций.
7	Форматы и первичная валидация сырых мультимодальных данных Рассматриваемые вопросы: - стандарты представления изображений, видео, аудио и телеметрии; - методы автоматической проверки целостности и метаданных файлов.
8	Инженерные методы очистки и синхронизации данных Рассматриваемые вопросы: - обработка пропусков, дубликатов и выбросов в сырых наборах; - синхронизация видеопотоков с логами бортовых систем по временным меткам.
9	Архитектура и развертывание серверов разметки Рассматриваемые вопросы: - компоненты и системные требования self-hosted инструментов cvat и label studio; - интеграция инструментов аннотации с корпоративной инфраструктурой.
10	Эргономика процесса аннотации и обработка пограничных случаев Рассматриваемые вопросы: - организация рабочего пространства разметчика и сценарии использования горячих клавиш; - стратегии фиксации и эскалации спорных и неоднозначных объектов.
11	Метрики оценки качества разметки и согласованности аннотаторов Рассматриваемые вопросы: - расчет intersection over union и метрик межэкспертной согласованности; - статистика коэна и анализ матрицы ошибок для многоклассовой разметки.
12	Статистический анализ датасета и выявление системных смещений Рассматриваемые вопросы: - методы визуализации распределения классов и атрибутов; - идентификация географических, временных и погодных смещений в выборке.
13	Стратегии балансировки классов и стратификации выборок Рассматриваемые вопросы: - методы передискретизации, недодискретизации и генерации синтетических данных; - стратифицированное разбиение датасета на обучающую и валидационную выборки.
14	Индустриальные форматы экспорта датасетов Рассматриваемые вопросы: - структурные особенности форматов coco, yolo и pascal voc; - требования к валидации геометрии и ссылочной целостности при экспорте.

№ п/п	Тематика лекционных занятий / краткое содержание
15	Версионирование данных и управление артефактами ML-проектов Рассматриваемые вопросы: - принципы работы data version control и отслеживания итераций датасетов; - интеграция dvc с системами контроля версий кода и удаленными хранилищами.
16	Паспортизация датасетов и оформление технической документации Рассматриваемые вопросы: - структура паспорта датасета и описание ограничений его использования; - требования государственных стандартов к оформлению пояснительных записок.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Анализ транспортной ML-задачи и выбор типа разметки Студент изучает бизнес-требования к выбранной транспортной задаче машинного обучения и анализирует предоставленный набор сырых данных. На основе анализа принимается обоснованное решение о выборе оптимального геометрического типа аннотации для целевых объектов. Результат выбора фиксируется в техническом дневнике проекта с приведением аргументов.
2	Первичная валидация и анализ сырых данных Студент получает набор сырых файлов и пишет скрипт для автоматической проверки их целостности и анализа метаданных. В процессе работы выявляются поврежденные или дублирующиеся записи, после чего формируется аналитический отчет о качестве исходного материала.
3	Проектирование иерархии классов и таксономии объектов Студент разрабатывает детальную схему классов для своей задачи, определяя атрибуты и правила взаимного исключения объектов. Разработанная таксономия оформляется в виде наглядной диаграммы и сводной таблицы с примерами для каждого выделенного класса.
4	Разработка гайдлайна разметчика Студент создает структуру технической инструкции в отечественном офисном пакете и пишет введение с описанием решаемой транспортной задачи. Формируется раздел с эталонными примерами аннотаций, содержащий визуальные иллюстрации и подробные пояснения к граничным случаям.
5	Очистка и синхронизация мультимодальных данных Студент разрабатывает программные скрипты для приведения данных к единому формату и удаления дубликатов. Выполняется синхронизация видеопотоков с логами бортовых систем транспорта по временным меткам для последующей комплексной разметки.
6	Развертывание и настройка инструмента разметки Студент устанавливает и настраивает сервер разметки в корпоративной операционной системе, создает новый проект и импортирует подготовленные данные. В интерфейсе инструмента настраивается схема классов в полном соответствии с ранее разработанной таксономией.
7	Выполнение разметки (пилотная выборка) Студент выполняет аннотацию пилотной выборки объектов строго по разработанному гайдлайну и фиксирует все спорные случаи. Анализируется эргономика процесса разметки и формулируются конкретные предложения по уточнению технической инструкции.
8	Уточнение гайдлайна и выполнение разметки (основной объем) Студент дорабатывает инструкцию на основе опыта пилотной разметки, добавляя раздел с разбором

№ п/п	Тематика практических занятий/краткое содержание
	сложных случаев. После утверждения финальной версии гайдлайна выполняется разметка основного объема данных с обязательной фиксацией всех отклонений.
9	Разметка сложных и пограничных случаев Студент целенаправленно размечает выборку сложных объектов, включая частично перекрытые и нестандартные ракурсы. Принятые решения документируются в отдельном разделе инструкции и формируется набор эталонных примеров.
10	Разработка скриптов автоматической валидации аннотаций Студент пишет программные скрипты для проверки геометрической корректности аннотаций и валидации структуры данных. Разработанные скрипты тестируются на пилотной выборке для выявления нулевых площадей, пересечений и выхода координат за границы изображений.
11	Расчет метрик качества разметки и анализ согласованности Студент применяет разработанные скрипты для расчета метрик межэкспертной согласованности и анализирует распределение ошибок по классам. Выявляются системные проблемы разметки и формируется аналитический отчет о качестве с визуализациями распределений.
12	Балансировка классов и фильтрация некачественных аннотаций Студент анализирует распределение классов в размеченном датасете и применяет методы стратифицированной выборки для устранения дисбаланса. Выполняется фильтрация аннотаций с низким качеством и формируется финальная сбалансированная выборка.
13	Экспорт датасета в индустриальные форматы Студент экспортирует готовый датасет в стандартный формат и проверяет корректность структуры выходных файлов. Проводится финальная валидация соответствия экспортированных аннотаций требованиям целевых фреймворков глубокого обучения.
14	Версионирование данных и кода Студент инициализирует репозиторий системы контроля версий и настраивает инструменты версионирования данных для отслеживания итераций датасета. Создается коммит с финальной версией данных, конфигурационными файлами и формируется описание структуры проекта.
15	Формирование паспорта датасета Студент заполняет паспорт датасета с описанием состава данных, ограничений использования и этических аспектов в отечественном офисном пакете. Описывается архитектура пайплайна подготовки и приводятся статистические характеристики итоговой выборки.
16	Оформление финальной документации к курсовому проекту Студент оформляет итоговую пояснительную записку по государственному стандарту с использованием отечественного офисного пакета. Готовится презентация и пакет артефактов для итоговой защиты курсового проекта.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендованной литературы.
2	Выполнение курсовой работы.
3	Подготовка к промежуточной аттестации.
4	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых работ

Проектирование таксономии и разработка гайдлайна для детекции дорожных знаков с экспортом в формат YOLO.

Подготовка и версионирование датасета для семантической сегментации дефектов дорожного покрытия.

Синхронизация видеопотоков и телеметрии бортовых систем для разметки событий экстренного торможения.

Создание технической инструкции и аннотирование выборки для распознавания государственных регистрационных знаков.

Развертывание self-hosted сервера CVAT и настройка пайплайна разметки для трекинга транспортных средств.

Подготовка сбалансированного датасета для instance-сегментации повреждений транспортных средств.

Проектирование иерархии классов и разметка данных для системы мониторинга состояния водителя.

Очистка, валидация и аннотирование телеметрии сельскохозяйственной техники для задач автономного вождения.

Разработка скриптов автоматической валидации геометрии аннотаций и контроля качества разметки пешеходов.

Формирование и паспортизация датасета для задачи подсчета пассажиров в салоне общественного транспорта.

Подготовка мультимодального датасета для контроля габаритов приближения на железнодорожном транспорте.

Стратификация и балансировка классов при создании датасета для детекции микромобильных транспортных средств.

Аннотирование временных рядов телеметрии и выявление аномалий в логах бортовых систем грузового транспорта.

Проектирование воспроизводимого пайплайна от сырых мультимодальных данных до версионированного датасета в формате COCO.

Комплексная подготовка, разметка и оформление по ГОСТ паспорта датасета для системы распознавания дорожной разметки.

5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

№ п/п	Библиографическое описание	Место доступа
1	Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721 (дата обращения: 22.06.2026)

2	Инженерия данных в Python : руководство / пер. с англ. А. Ю. Гинько. — Москва : ДМК Пресс, 2025. — 528 с. — ISBN 978-5-93700-381-2. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/514945 (дата обращения: 22.06.2026)
3	Конкина, В. В. Введение в большие данные и анализ информации : учебное пособие / В. В. Конкина, А. Б. Борисенко, И. Л. Коробова. — Тамбов : ТГТУ, 2024. — 82 с. — ISBN 978-5-8265-2749-8. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/472331 (дата обращения: 22.06.2026)
4	Уржумов, Д. В. Системы распознавания образов. Компьютерное зрение: практикум : учебное пособие / Д. В. Уржумов, А. В. Кревецкий. — Йошкар-Ола : ПГТУ, 2024. — 36 с. — ISBN 978-5-8158-2386-0. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/437471 (дата обращения: 22.06.2026)
5	Кэлер, А. Изучаем OpenCV 3. Разработка программ компьютерного зрения на C++ с применением библиотеки OpenCV / А. Кэлер, Г. Брэдски ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2017. — 826 с. — ISBN 978-5-97060-471-7. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/108126 (дата обращения: 22.06.2026)
6	Колмогорова, С. С. Основы искусственного интеллекта : учебное пособие для студентов / С. С. Колмогорова. — Санкт-Петербург : СПбГЛТУ, 2022. — 108 с. — ISBN 978-5-9239-1308-8. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/257804 (дата обращения: 22.06.2026)
7	Методы искусственного интеллекта в обработке данных и изображений : монография / А. Ю. Дёмин, А. К. Стоянов, В. Б. Немировский, В. А. Дорофеев. — Томск : ТПУ, 2016. — 130 с. — Текст : электронный	Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/106257 (дата обращения: 22.06.2026)

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

Электронно-библиотечная система «Лань» – URL: https://e.lanbook.com/.

Официальная документация библиотеки Pandas – URL: https://pandas.pydata.org/docs/.

Официальная документация библиотеки OpenCV – URL: https://docs.opencv.org/.

Официальная документация инструмента CVAT – URL: https://docs.cvat.ai/.

Официальная документация инструмента Label Studio – URL: https://labelstud.io/guide/get_started.

Официальная документация DVC (Data Version Control) – URL: https://dvc.org/doc.

ГОСТ 7.32-2017. Отчет о научно-исследовательской работе. Структура и правила оформления – URL: https://docs.cntd.ru/document/1200147998.

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Операционные системы – Astra Linux / ALT Linux / РЕД ОС.

Офисные пакеты – Р7-Офис / МойОфис (для подготовки отчетов и презентаций по ГОСТ).

Среда разработки – Anaconda Distribution, Jupyter Notebook / JupyterLab, VS Code Community Edition (оффлайн-версии).

Технологический стек ИИ и Data Science – Python 3.10+, Pandas, NumPy, OpenCV, Matplotlib, Seaborn.

Разметка данных – CVAT / Label Studio (развертывание в локальном контуре через Docker).

Версионирование данных и кода – Git, DVC (Data Version Control).

Работа с телеметрией – стандартные парсеры CSV/JSON, библиотеки для синхронизации временных рядов.

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю).

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

Для практических занятий – наличие персональных компьютеров вычислительного класса.

9. Форма промежуточной аттестации:

Зачет в 5 семестре.

Курсовая работа в 5 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).

Авторы:

старший преподаватель кафедры
«Цифровые технологии управления
транспортными процессами» А.Ю. Кремнев

Согласовано:

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической
комиссии Н.А. Андриянова