УДК 004.8
doi:10.31799/1684-8853-2026-3-35-48
EDN: XBAJEX

Научные статьи
Articles

Метод обучения моделей компьютерного зрения на основе кросс-модальной дистилляции знаний с применением больших визуальных моделей

А. В. Кучков[а], *аспирант, orcid.org/0009-0007-7508-3348*
А. М. Кашевник[а,б], *канд. техн. наук, доцент, orcid.org/0000-0001-6503-1447, alexey.kashevnik@iias.spb.su*
[а]*Университет ИТМО, Кронверкский пр., 49, Санкт-Петербург, 197101, РФ*
[б]*Санкт-Петербургский институт информатики и автоматизации РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *имеющиеся методы обучения мультимодальных моделей компьютерного зрения в большинстве случаев представляют собой отдельные ветви выделения распознавания признаков с поздним смешением результатов.* ***Цель:*** *разработать метод создания мультимодальных моделей компьютерного зрения с использованием единого представления мультимодальных данных для упрощения процессов смешения данных и дистилляции знаний.* ***Методы:*** *сериализация разреженных и плотных типов данных; кросс-модальная дистилляция знаний для архитектур компьютерного зрения; применение больших визуальных моделей для дистилляции знаний в сериализованном формате.* ***Результаты:*** *разработан метод обучения моделей компьютерного зрения на основе кривых Пеано с использованием дистилляции знаний из больших визуальных моделей. Метод позволяет производить смешение данных различных размерностей с помощью кросс-модального внимания в реальном времени посредством применения одномерных кривых Пеано (кривых Гильберта и Мортона) для сериализации многомерных данных. Предложенный метод показал задержку 50 мс против 40 мс в одномодальном режиме (Point Transformer v3), что свидетельствует о низких накладных расходах при кросс-модальной дистилляции на сериализованных картах признаков. Метод протестирован в режиме предобучения на датасете nuScenes с обращением к большой визуальной модели DINOv3. В режиме дистилляции использование 25 % от общего набора данных обеспечило 79,2 mIoU по сравнению с 82,1 mIoU при 100 % набора данных с функцией потерь — коэффициентом Отиаи.* ***Практическая значимость:*** *с использованием сериализованного представления данных методы кросс-модального смешения станут менее ресурсозатратными.* ***Обсуждение:*** *предложенный метод позволяет унифицировать декодер в модели сегментации смешанных данных благодаря кросс-модальному смешению сериализованных признаков после энкодеров изображений и облаков точек. При этом ранняя сериализация изображений показала себя нецелесообразной ввиду изначально плотной структуры изображений. Реализация метода сериализации изображений с меньшим временем выполнения даст возможность отказаться от раздельных энкодеров для ветвей облаков точек и изображений, что может существенно упростить архитектуру.*

Ключевые слова *— мультимодальные модели, дистилляция знаний, большие визуальные модели, кривые Пеано, кросс-модальное внимание.*

Для цитирования: Кучков А. В., Кашевник А. М. Метод обучения моделей компьютерного зрения на основе кросс-модальной дистилляции знаний с применением больших визуальных моделей. *Информационно-управляющие системы*, 2026, № 3, с. 35–48. doi:10.31799/1684-8853-2026-3-35-48, EDN: XBAJEX
For citation: Kuchkov A. V., Kashevnik A. M. Method for training computer vision models based on cross-modal knowledge distillation using large visual models. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 35–48 (In Russian). doi:10.31799/1684-8853-2026-3-35-48, EDN: XBAJEX

Введение

Мультимодальные модели компьютерного зрения предполагают использование двух и более сенсоров машинного зрения для восприятия окружающей обстановки в целях реагирования на изменения во внешней среде. Примерами подобных средств являются системы автономного вождения и системы-ассистенты предупреждения о столкновениях. Мультимодальные модели ввиду большой емкости требуют большого количества данных для обучения. Настоящая проблема стала еще более актуальной с появлением сетей-трансформеров, демонстрирующих высокие показатели точности, но требующих больше данных для их обучения. Для обучения мультимодальных моделей используются датасеты, содержащие данные от нескольких сенсоров, таких как лидары, камеры, радары и иные датчики. Среди публично доступных датасетов наибольшим объемом обладают такие наборы данных, как KITTI, nuScenes, SemanticKITTI и Waymo Open Dataset. Независимо от обилия датасетов количество взаимно аннотированных данных для нескольких сенсоров в рамках одной сцены ограничено. Несмотря на то, что датасеты KITTI, nuScenes и SemanticKITTI представляют облака точек и изображения для каждой из сцен, приоритетным сенсором остается лидар, данные которого, как правило, являются основным источником аннотаций.

Метки, явно заданные в датасете и содержащие в себе информацию о классе, форме и положении объекта, являются «жесткими» метками, количество которых зачастую ограниченно для определенных типов данных.

В общем случае подходы к уменьшению зависимости от «жестких» меток можно разделить на четыре категории [1, 2]:

— weakly-supervised-методы — использование меток, сгенерированных другой моделью-учителем («мягких» меток) [3];

— semi-supervised-методы — совместное использование небольшого размеченного набора данных и большого неразмеченного набора [4];

— self-supervised-методы — обучение с генерацией псевдометок самой моделью без привлечения внешней разметки [5];

— unsupervised-методы — полный отказ от использования как «жестких», так и «мягких» меток в пользу иных критериев оценки качества предсказаний [6].

Модель учителя представляет собой сеть, из которой происходит дистилляция знаний. Модель студента (ученика) при этом является сетью, в которую дистиллируются данные [7].

Weakly-supervised-методы используют метки, сгенерированные иной моделью и называемые «мягкими» метками. Качество сетей, обученных weakly-supervsied-методом, зависит как от модели учителя, так и от используемой функции ошибки дистилляции. Как правило, модель учителя обучена на большем наборе данных по сравнению с сетью ученика. Преимуществом метода является то, что отсутствуют какие-либо ограничения на архитектуру модели — ученика, что позволяет использовать как устоявшиеся подходы (CNN), так и более новые архитектуры, такие как визуальные трансформеры (ViT) [3].

Self-supervised-методы обычно реализуют архитектуру с двумя ветвями, в которой одна ветвь является моделью учителя, а другая — моделью студента. Их архитектуры могут как быть полностью идентичными, так и иметь некоторые особенности. Преимуществом данного метода является то, что он позволяет произвести обучение в один этап, а также отлично масштабируется вплоть до моделей с 7 млрд параметров, что допускает создание таких больших визуальных моделей (VFM, Visual Foundation Models), как DINOv3 (Distillation with No Labels) [5], SAM (Segment Anything Mode) [4], SegGPT (Segmenting Everything In Context) [8]. Однако в то время, как данный метод очень хорошо подходит для обучения больших моделей, в том числе визуальных, на текущий момент он не адаптирован под задачи реального времени, что особенно критично при выполнении нейронных сетей на встраиваемом оборудовании с ограниченными ресурсами.

Unsupervised-методы в свою очередь подразумевают полное отсутствие контроля над обучением посредством каких-либо меток. В классическом понимании unsupervise-обучение применяется для кластеризации, оценки плотности вероятности и снижения размерности пространства признаков.

В настоящей работе основное внимание уделяется weakly-supervised-методу со сжатием признаков модели учителя в пространство меньшей размерности сети ученика. Преимущество указанного подхода заключается в отсутствии необходимости применять «жесткие» метки для вычисления критериев точности работы моделей, так как на этапе предобучения происходит дистилляция знаний из cls-токена сети VFM в сеть PTv3, что позволяет ускорить процесс обучения. Таким образом, сеть DINOv3 выступает в качестве генератора self-supervised признаков для нашей модели.

Целью данной работы является создание метода обучения мультимодальных моделей, использующих изображения и облака точек, на основе сериализации данных и привлечения VFM – DINOv3. Научная значимость настоящего исследования заключается в следующем:

1) предложен метод совместного представления признаков облаков точек изображений в сериализованном виде с помощью кривых Пеано;

2) разработан метод смешения признаков изображений и облаков точек в сериализованном пространстве;

3) разработан метод дистилляции знаний в сериализованное пространство признаков с использованием VFM – DINOv3.

Обзор исследований обучения моделей компьютерного зрения на основе дистилляции знаний

Непосредственный метод дистилляции знаний зависит от модальности дистиллируемых данных. Методы дистилляции знаний для различных модальностей представлены в табл. 1. Дистилляция знаний внутри одной модальности, как правило, осуществляется в целях сжатия знаний исходной модели в веса меньшей.

Наибольший интерес представляют методы переноса знаний в направлениях Image → Lidar и Image → Image ввиду разнообразия таких VFM, как DINOv3, SAM и SegGPT, обученных self-supervsied-методом на больших наборах данных. VFM могут служить в качестве учителя для генерации качественных карт признаков без необходимости дообучения. Вследствие большой емкости указанных моделей область их применения не ограничена лишь RGB-изображениями; они могут

■ ***Таблица 1.*** Методы кросс-модальной дистилляции знаний
■ ***Table 1.*** Cross-modal knowledge distillation methods

Данные	Типы дистилляции знаний	Примеры	Функции потерь	Применение
Изображения и облака точек	Выравнивание признаков, кросс-внимание, контрастная дистилляция знаний	DITR [9], Modal Mimicking [10], OLIVINE [3]	Функция ошибки InfoNCE между сетями 2D и 3D, коэффициент Отиаи, расстояние Кульбака – Лейблера	Предобучение 3D-моделей при помощи 2D-учителей
Облака точек радара и лидара	Выравнивание признаков, кросс-внимание, контрастная дистилляция знаний	RadarDistill [11]	InfoNCE, коэффициент Отиаи, расстояние Кульбака – Лейблера	Предобучение радарных моделей при помощи облаков 3D-учителей
Текст и облака точек	Дистилляция с учетом интра- и кросс-модальных отношений	Multimodal Relation Distillation [12], CLIP2Point [13]	NT-Xent Loss, коэффициент Отиаи, расстояние Джефри, нормализованное сходство	Предобучение 3D-моделей с LLM (большой языковой моделью)
Текст и изображения	Контрастная дистилляция знаний	CLIP [14], TinyCLIP [15], SCKD [16]	InfoNCE, MSE	Предобучение 2D-моделей с LLM

■ ***Таблица 2.*** Большие визуальные модели
■ ***Table 2.*** Visual Foundation Models

Модель	Модальность	Параметры	Задачи
DALL-E [17]	Текст + изображения	~12 млрд (v1), 3,5 млрд (v2)	Генерация изображений
SAM [4]	Текст + изображения	До ~636 млн	Сегментация изображений
Stable Diffusion [18]	Текст + изображения	До ~8 млрд	Генерация изображений
CLIP ViT-L/14 [14]	Текст + изображения	~ 428 млн	Сопоставление изображений и текста
Florence-2 [19]	Текст + изображения	230 млн (Base), 770 млн (Large)	Детектирование, сегментация
DINOv3 [5]	Изображения	До ~7 млрд	Извлечение универсальных визуальных признаков
CogVLM-17B [20]	Текст + изображения	17 млрд	Визуальные ответы на вопросы
OpenVLA [21]	Текст + изображения	7 млрд	Генерация действий для манипуляций
Qwen2.5-VL-32B-Instruct [22]	Текст + изображения	32 млрд	Мультимодальные задачи (текст + изображение)
Gemma 3 [6]	Текст + изображения	1/4/12/27 млрд	Мультиязычная VLM

применяться для задач дистилляции знаний и для модальностей, отличных от RGB-изображений, таких как лидарные и радарные данные.

Поскольку обучающий набор не обязан содержать истинные метки, то для процесса дистилляции могут использоваться неразмеченные экспериментальные данные с последующим получением «слабых» меток с помощью VFM. Использование VFM зачастую подразумевает генерацию «слабых» меток до начала процесса предобучения. Выбор VFM определяется исходя из требований к емкости модели, точности, а также ее доступности. В табл. 2 представлены актуальные на текущий момент VFM.

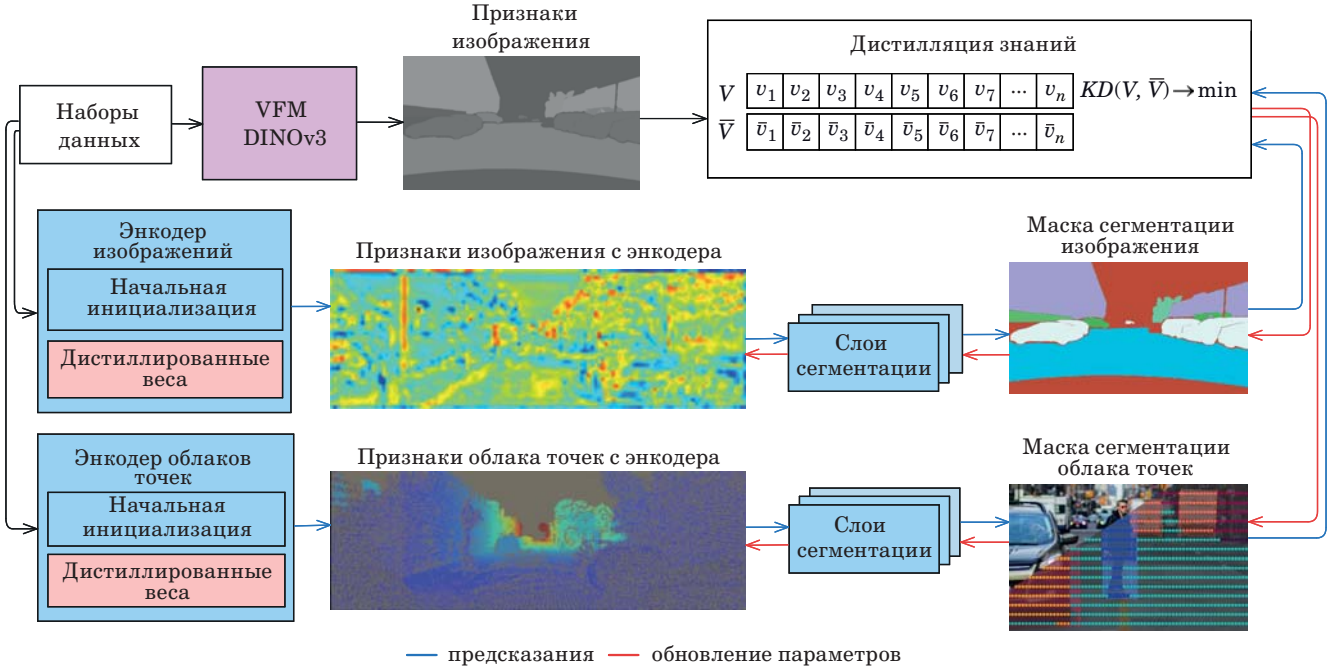
Как правило, для получения «слабых» меток используется специальный тип моделей VFM, обученных на выполнение общих задач компьютерного зрения (сегментация, детектирование). В данную группу входят такие модели, как DINOv1/2/3, SAM, CLIP и Qwen2.5-VL.

Предобучение мультимодальных моделей с помощью VFM подразумевает индивидуальный подход к дистилляции знаний для данных каждого из сенсоров. Более того, дистилляция может осложняться применением смешения данных в модели. Для дистилляции знаний в подобных моделях необходимо решить несколько вопросов: на каком этапе модели происходит дистилляция, из какой модальности дистиллируются данные и какую функцию ошибки дистилляции использовать?

Этап дистилляции, как правило, происходит в поздних слоях сети ввиду наличия качественных карт признаков [3, 9]. Дистилляция на ранних этапах малоэффективна, так как первые слои не содержат глубоких признаков. Модальность дистилляции определяется наличием VFM, обученной на большом объеме данных. В связи с тем, что большинство современных VFM обучены на RGB-изображениях, использование данной модальности на текущий момент является общепринятым подходом.

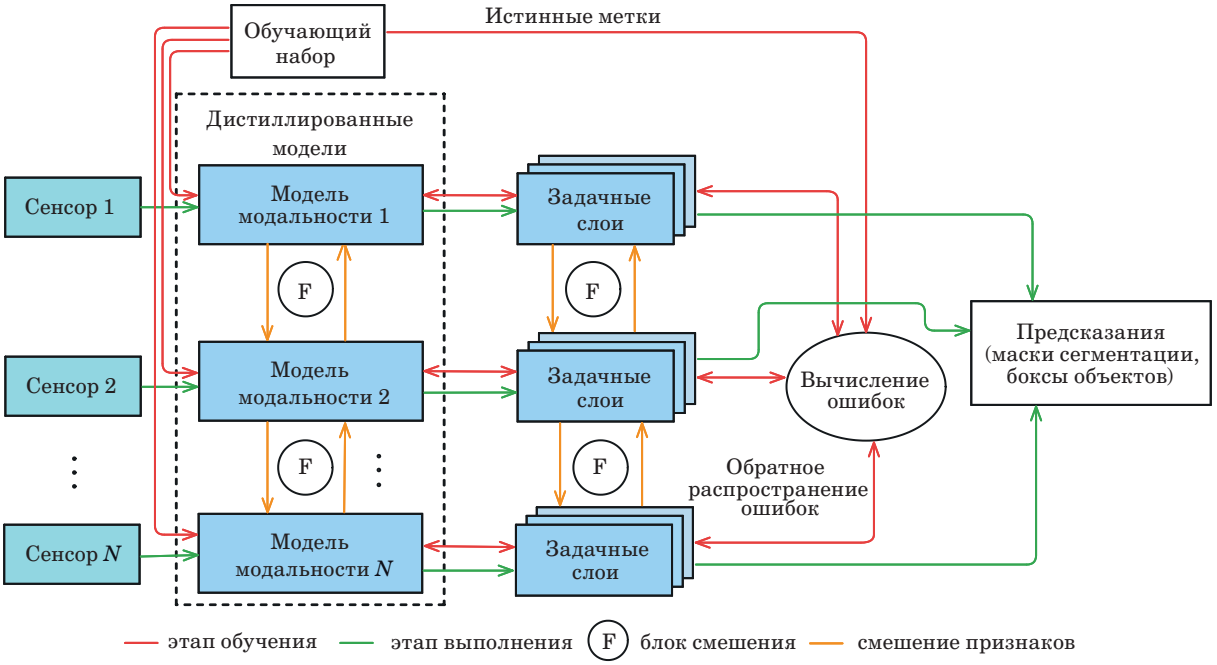
Для выбора функции ошибки дистилляции необходимо учитывать, что не все функции ошибок позволяют оценить локальные и глобальные признаки. В связи с чем для дистилляции, как правило, применяются сразу несколько функций ошибок. Схема обучения двухмодальной модели с применением механизма дистилляции знаний показана на рис. 1. Применение подобных моделей охватывает три этапа: предобучение, обучение и выполнение. Дистилляция знаний происходит исключительно на первом этапе, который занимает большую часть времени. Источником данных на стадиях предобучения и обучения служат такие наборы данных, как KITTI, nuScenes, Waymo Open Dataset и др. 2D- и 3D-модели ответственны за выделение признаков из облаков точек и изображений. VFM в свою очередь (DINOv3 на схеме) выделяет признаки из изображений. Дистилляция происходит на двух масштабах: на семантическом уровне и на уровне признаков.

Далее дистиллированные модели проходят через этап дообучения (fine-tuning), где используются истинные метки из обучающего набора для необходимых модальностей. По причине того, что модель уже является предобученной, количество данных в обучающем наборе может быть сокращено вплоть до 1 % [3]. Более того, возможен сценарий Linear Probing (LP), в котором замораживается сеть выделения признаков, а обучению подвергаются лишь задачно-ориентированные слои (рис. 2). Данный метод позволяет дообучить модель под конкретную задачу гораздо быстрее, так как требуется не полное обновление параметров модели, а лишь нескольких слоев. Сценарий LP может быть применен к произвольному количеству источников данных.

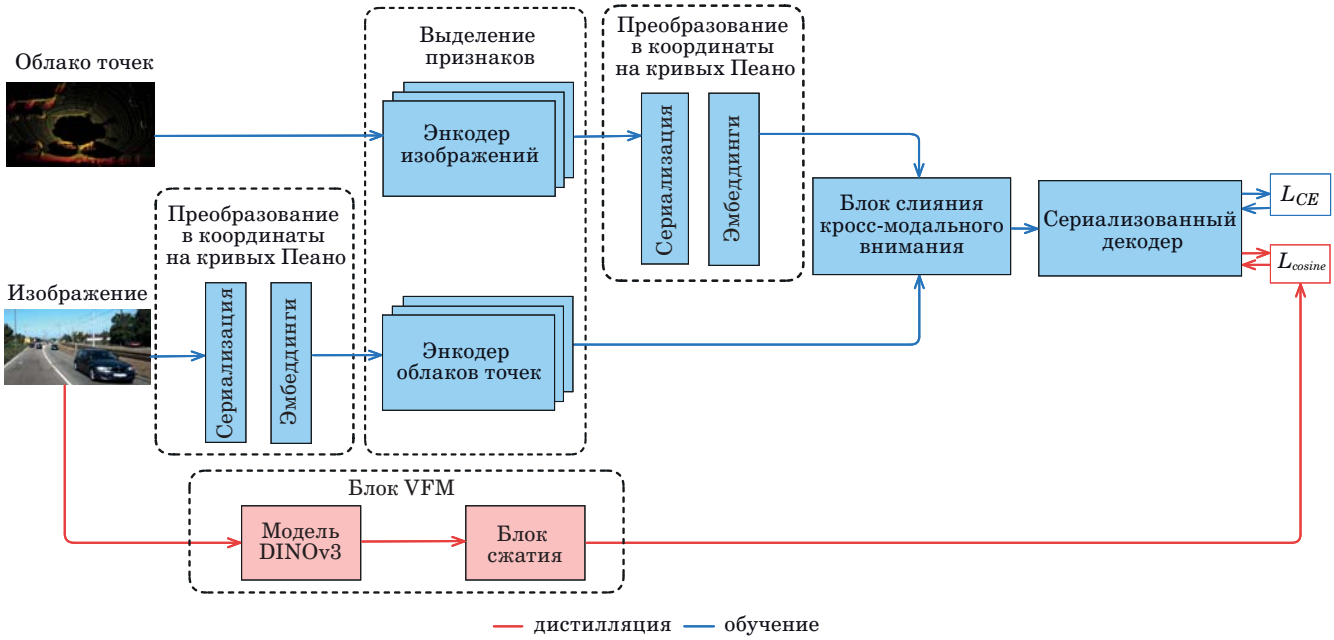


■ ***Рис. 1.*** Пример обучения моделей обработки облаков точек и RGB-изображений с предложенным методом дистилляции знаний

■ ***Fig. 1.*** An example of multimodal model pretraining with LiDAR and RGB input using the proposed knowledge distillation method



■ ***Рис. 2.*** Смешение данных в мультисенсорной системе с дистилляцией знаний
■ ***Fig. 2.*** Data fusion in a multisensory system with knowledge distillation



■ ***Рис. 3.*** Обучение моделей компьютерного зрения на основе дистилляции знаний с использованием VFM
■ ***Fig. 3.*** Training computer vision models based on knowledge distillation using VFM

Метод дистилляции знаний в мультисенсорной модели

В настоящем разделе предлагается метод создания и обучения мультимодальных моделей с использованием кривых Пеано и дистилляции знаний с применением VFM. Метод предполагает наличие следующих компонентов: источника данных (наборов данных для обучения), сетей выделения признаков для каждого сенсора (энкодеров), блока VFM, блока смешения данных на основе кросс-модальной дистилляции, а также энкодера смешанных признаков (рис. 3). Сначала облака точек подвергаются процессу сериализации,

который трансформирует 3D-координаты в координаты на одномерной кривой Пеано, проходящей через точки координатной системы с заданным размером сетки. Данному процессу подвергается лишь облако точек ввиду его разреженной структуры. Затем происходит этап выделения признаков, на котором из облака точек и изображений формируются карты признаков. Дальнейшим шагом является преобразование всех карт признаков в один формат — координаты на кривых Пеано. Для этого признаки изображений сериализируются и разбиваются на части, формируя понятную для декодера-трансформера последовательность. Перед декодером стоит блок смешения данных на основе кросс-модального внимания, преобразующий две одномерные последовательности на входе в одну с помощью двух блоков внимания.

Сериализация позволяет сократить расходы на вычисление матриц внимания путем устранения необходимости построения больших деревьев для вычисления взаимного расположения элементов облаков точек и изображений. Смешанные данные попадают в декодер, структура которого является зеркальной к энкодеру. Декодер на выходе формирует классы для каждой точки из облака.

Во время обучения в качестве основного критерия используется ошибка кросс-энтропии (Cross Entropy, CE), в то время как для предобучения используется коэффициент Отиаи.

Сериализация карт признаков

Для представления признаков облаков точек и изображений в едином пространстве используется техника сериализации многомерных данных [23]. Сериализация осуществляется с помощью одной или нескольких кривых Пеано, преобразующих облака точек и признаки изображений в набор координат. Обучение моделей компьютерного зрения на основе дистилляции знаний с использованием VFM одномерной кривой:

$$\varphi^{-1} : Z^n \to Z^1,$$

где φ^{-1} — функция отображения n-мерной точки в определенное положение на линии.

Необходимость в сериализации облака точек обусловлена неупорядоченной природой облаков точек, которая мешает применению механизмов внимания, использующихся в сетях-трансформерах. Следующая причина необходимости в иной форме представления данных возникает из-за того, что мультимодальные модели компьютерного зрения обладают высокой вычислительной сложностью ввиду высокой размерности данных. Как показано в работе [23], комбинация различных механизмов сериализации улучшает производительность моделей.

В случае с изображениями сериализация также может иметь место, что показано в пилотном исследовании, однако вычисление признаков на одномерной кривой не имеет никаких преимуществ по сравнению со стандартными методами выделения признаков, так как изображения по своей природе уже упорядочены.

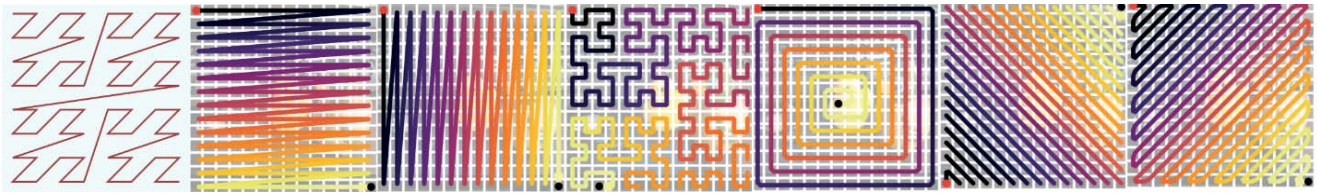
В настоящей работе делается акцент на том, что использование кривых Пеано позволяет реализовать механизм кросс-модального внимания с последующим смешением данных более эффективно, чем это можно было бы сделать с помощью вокселизации или проецирования облаков точек на 2D-плоскость. Альтернативным подходом является использование поточечной (pointwise) свертки для ускорения вычисления слоя самовнимания (self-attention) [24]. Однако данный подход слабо применим для трехмерных признаков облака точек.

Сериализованное представление многомерных координат на кривой характеризуется дискретным шагом g, представляющим размер сетки. В текущей работе подобной трансформации подвергаются не только облака точек, но и признаки изображения, т. е. для каждого признака ищется соответствие на кривой Пеано. Также, следуя работе [23], для кодирования батчей используются k нулевых старших разрядов позиции пикселя:

$$Encode(\mathbf{p}, b, g) = (b << k) \mid \varphi^{-1}(|\mathbf{p}/g|),$$

где $\mathbf{p}$ — координаты пикселя; b — размер батча; g — размер ячейки; << — битовый сдвиг влево.

Для сохранения возможности обратного преобразования, а также для ускорения вычислений вместо трансформации облаков и изображений непосредственно в формат кривых Пеано используются лишь индексы, ставящие соответствие каждой точке или пикселю координату на кривой. В работе [25] авторы выделили наиболее подходящие типы кривых (рис. 4). В настоящей работе для сериализации используются кривые Мортона и Гильберта. Обе кривые обладают свойством сохранять взаимное расположение двух точек, т. е. индексы точек, лежащих близко в N-мерном пространстве, также будут лежать близко друг к другу. При этом кривая Гильберта обладает данным свойством в большей степени, чем кривая Мортона. Остальные типы пространственных кривых таким свойством не обладают и, соответственно, не подходят для вычисления self-attention на признаках, полученных с облаков точек, ввиду наличия разрывов при переходе через строки или столбцы входного тензора. В работе [25] авторы показали, что в зависимости от датасета и используемой модели форма оптимальной аппроксимирующей кривой может



■ ***Рис. 4.*** Кривые Пеано. Слева направо: кривая Мортона (z-order), разбиение по строкам, разбиение по столбцам, кривая Гильберта, спираль, диагональ, змеевидная кривая

■ ***Fig. 4.*** Peano curves. From left to the right: Morton (z-order), row order, column order, Hilbert, spiral, diagonal, snake-like curves

меняться. Учитывая данную информацию, для сравнения точности модели в текущей работе были протестированы как отдельные кривые Мортона и Гильберта, так и их комбинации.

Важным свойством любой кривой является сохранение пространственной близости точек, имеющих близкие координаты. Не все кривые обладают данным свойством и имеют периодические разрывы. Так, на кривых, построенных методом разбиения по строкам и столбцам, отмечаются разрывы через каждые N элементов.

Дистилляция сериализованных данных

На этапе дистилляции применяется метод, описанный в DITR (DINO in the Room) [9]. Данный подход заключается в вычислении коэффициента Отиаи в качестве функции ошибки для направления процесса дистилляции в сторону схожести карт признаков PTv3 и DINOv3. Для дистилляции из DINOv3 были взяты карты признаков из iBOT [5] подсети, так как данные признаки содержат информацию на уровне патчей изображения. Таким образом, минимизировалась следующая ошибка:

$$L_{cosine} = \sum [1 - x_i^{pred}\, x_i^{2D} \div (||x_i^{pred}||\ \ ||x_i^{2D}||)] \div \upsilon,$$

где υ — набор точек, видимый как минимум на одной камере.

Смешение данных

Наиболее распространенным подходом к смешению данных является раздельная обработка данных различных сенсоров ввиду простоты адаптации существующих решений [26]. Авторы [27] подтвердили, что такие методы смешения мультимодальных данных, как конкатенация признаков и сложение признаков, хоть и могут работать при определенных сценариях, не дают существенного превосходства над одномодальной сетью из-за слабой корреляции признаков. При этом в последнее время все чаще применяются механизмы смешения данных на основе кросс-модального внимания [28, 29]. В настоящей работе для смешения данных также используется метод кросс-модального внимания с двумя Attention-ветвями с матрицами ключей **K** от модальностей изображений и облаков точек. Блок данной архитектуры показал свою эффективность в кросс-модальном смешении данных, включая геометрическую и визуальную последовательности [28]. Предварительно для избежания лишних вычислений облака точек фильтруются по полю зрения камер.

Блок кросс-модального смешения (рис. 5) представляет собой комбинирование двух блоков Multi-Head Attention MHA с вычислением произведения $\mathbf{QK}^{\mathrm{T}}$, где матрицы **Q** и **K** принадлежат разным модальностям. Вычисление Attention в данном случае является двунаправленным, так как используется два блока с вычислением коэффициентов схожести от лидара к камере и от камеры к лидару. Полученные таким образом карты признаков конкатенируются и подаются на вход декодера сети, который является общим для ветвей лидара и камеры. Получение смешанной карты признаков может быть описано следующим выражением:

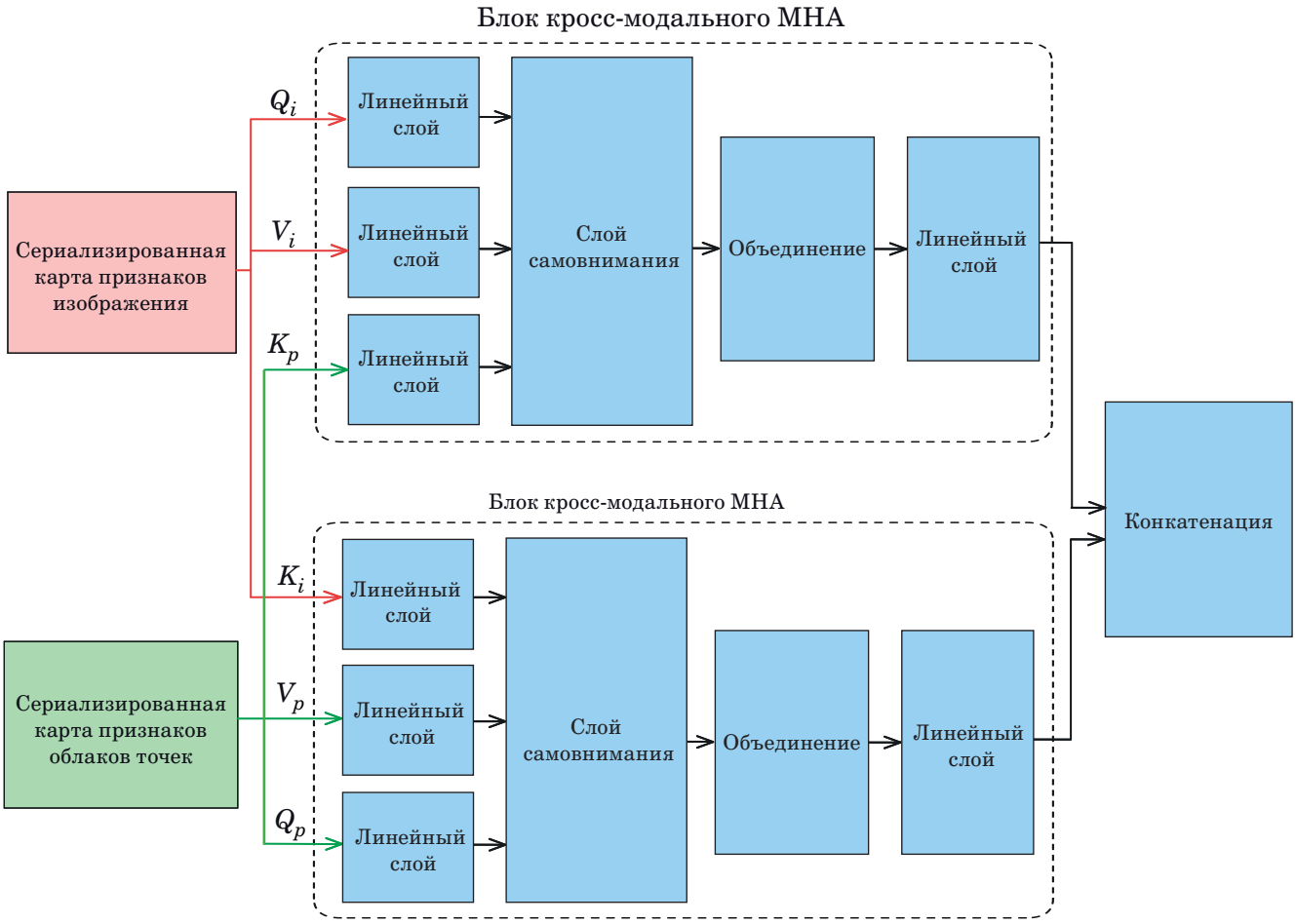
$$X_f = Concatenation(Attend_i(X_i, X_p), Attend_p(X_i, X_p)),$$

где $Attend_i$ и $Attend_p$ — внимание между изображениями (X_i) и облаками точек (X_p).

Эксперименты

В качестве базовой модели в текущей работе принята сеть 3D-трансформер PTv3, являющаяся SOTA-моделью на датасете nuScenes. Для дистилляции знаний применялась большая визуальная модель DINOv3 с ее функцией ошибки iBOT, позволяющей получить признаки на уровне входных частей изображений. Для сериализации облаков точек и изображений использован блок сериализации, преобразующий исходные облака точек и признаки изображений в координаты на кривой Пеано.

Эксперимент поставлен следующим образом. В первой части проводилось пилотное исследование на предмет возможности применения сети PTv3 для выделения признаков из изображений.

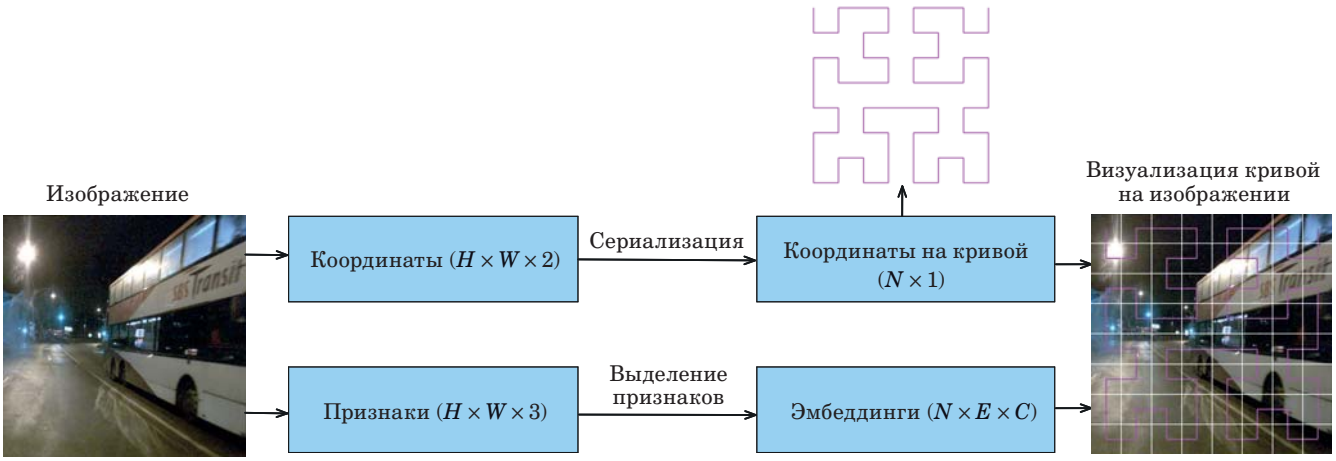


■ ***Рис. 5.*** Кросс-модальное смешение с помощью метода кросс-внимания
■ ***Fig. 5.*** Cross-modal data fusion with cross-modality attention

Во второй части исследовалась применимость механизма дистилляции знаний для обучения PTv3 с одномодальным мультимодальным входом.

Для проверки возможности выделения признаков из изображений был взят датасет CIFAR-10 [30], состоящий из 60 000 цветных изображений размером 32 × 32 с 10 классами, по 6000 изображений на каждый класс. Датасет поделен на 50 000 изображений в тренировочном наборе и 10 000 изображений в тестовом наборе. Из сети PTv3 был убран декодер, вместо которого использованы полносвязные слои для классификации изображений. Для каждого изображения построены координаты пикселей и каналы признаков (цвета RGB). Процесс сериализации продемонстрирован на рис. 6. Исследованы следующие варианты сериализации: сериализация кривыми Гильберта, Мортона, инвертированной кривой Гильберта, инвертированной кривой Мортона, а также сериализация с последовательным применением данных кривых.

В качестве каналов, содержащих признаки данных, использованы интенсивности изображения вместо интенсивности облака точек. Все вычисления выполнялись на NVIDIA RTX 3080. Размер набора данных — 32, количество эпох — 25, размер изображения — 32 × 32. Для лучшей визуализации координаты кривой отложены поверх оригинального изображения. В сравнение были включены различные комбинации из кривых Гильберта и Мортона, а также из их транспонированных вариантов. Результаты обучения трансформера PTv3 на изображениях представлены в табл. 3. Проведено сравнение метрик качества сверточных сетей VGG16 [31], ResNet50 [32] и Resnet18 [32] (которые зачастую используются как базовые для извлечения признаков) с результатами работы PTv3 при различных методах сериализации. Результаты пилотного исследования показали недостаточную эффективность метода сериализации для выделения признаков с изображения ввиду низкой скорости выполнения.



■ ***Рис. 6.*** Последовательность действий сериализации изображения
■ ***Fig. 6.*** Image serialization process. For visualization purposes, the curve is drawn over the original image

В то же время этот метод может применяться для смешения данных уже после процесса выделения признаков. По итогам первой части экспериментов на текущий момент было решено отказаться от PTv3 энкодера для ветви изображений. Для выделения признаков с изображений выбрана сеть Resnet50, потому как она обладает компромиссными характеристиками по скорости работы и точности. При сериализации признаков применялась комбинация кривых Мортона, Гильберта и их транспонированных вариантов, поскольку она продемонстрировала наивысшие показатели точности классификации как в пилотном исследовании, так и в самой работе PTv3.

■ ***Таблица 3.*** Метрики PTv3 с модальностью изображений и сравнение с другими моделями
■ ***Table 3.*** PTv3 metrics trained with image modality and comparison with other models

Модель	Точность, %	Параметры	Задержка, мс
PTv3 ($H + H_{trans} + Z + Z_{trans}$)	0,989	25,6 M	~31
PTv3 (H)	0,859	25,6 M	~28
PTv3 ($H + H_{trans}$)	0,890	25,6 M	~28
PTv3 ($H + Z$)	0,883	25,6 M	~28
PTv3 (Z)	0,969	25,6 M	~26
PTv3 ($Z + Z_{trans}$)	0,860	25,6 M	~26
VGG16 (32×32)	0,933	138 M	~9,9
ResNet18 (32×32)	0,813	11,5 M	~6,4
ResNet50 (32×32)	**0,922**	**~25 M**	**~8,3**

Примечание: H, Z и H_{trans}, Z_{trans} — кривые Гильберта, Мортона и их транспонированные варианты соответственно.

Облака точек подвергались сериализации с последующим выделением признаков с помощью сериализованного энкодера сети PTv3. Дальнейшее смешение данных происходило уже в сериализованном виде. Сравнение проводилось с предыдущими версиями Point Transformer, а также с моделями, основанными на схожем подходе. DITR [9] использует PTv3 и DINOv2 для генерации мультимодальных карт признаков и производит многоступенчатое смешение данных в декодере модели. Transformer Based Lidar Camera Fusion [33] использует Multi-Head Self-Attention на смешанной последовательности для повышения точности сегментации. LidarFormer [34] внедрил Cross-Space transformer для извлечения глобальной информации из BEV-карты признаков, а также Cross-Task transformer для извлечения семантической и объектной информации из смешанной карты признаков. Для предлагаемой в настоящей работе модели обучение производилось в двух сценариях (табл. 4): Linear Probing и с дообучением на частях данных из обучающего датасета nuScenes. В сценарии LP энкодер модели оставался замороженным, в то время как декодер подвергался обучению. Данные приведены для валидационной (val) и тренировочной (train) выборок.

Для каждой модели были произведены измерения времени выполнения одного батча на 16 элементов. Время задержки, приведенное в последнем столбце, является полным временем выполнения

■ ***Таблица 4.*** Результаты сравнительного анализа
■ ***Table 4.*** Results of the comparative analysis

Модель	LP		Дообучение на количестве данных, %						Веса	Время, мс
			1	5	10	25	100			
	val	train					val	train		
DITR [9]	–	–	–	–	–	–	**84,2**	**85,1**	7 B+25,6 M	~800
PTv3 [23]	–	–	–	–	–	–	81,2	83,0	25,6 M	**~ 40**
BEVFusion [35]	–	–	–	–	–	–	62,7	–	–	~120
T, B, LiDAR-Camera Fusion [33]	–	–	–	–	–	–	80,6	–	–	~83
LidarFormer [34]	–	–	–	–	–	–	82,7	81,5	77 M	~530
PTv3 Fusion	57,4	58,2	60,1	65,3	70,1	79,2	82,1	84,1	31,1 M	~50

(включая процесс сериализации для моделей PTv3 и PTv3 Fusion и копирования данных из CPU в GPU и обратно). Из представленных моделей DITR обладает наибольшим mIoU как на валидационном, так и на тренировочном датасете. При этом DITR также имеет самое большое время выполнения ввиду использования VFM DINOv2. LidarFormer, занимая второе место, обладает меньшей задержкой в 530 мс, хотя и недостаточной для задач реального времени.

Представленная в настоящей работе модель PTv3 Fusion превосходит оригинальную модель по точности на валидационном датасете (82,1 против 81,2 у PTv3). При этом, в отличие от DITR, она обладает временем, достаточным для выполнения задач реального времени. Дополнительные накладные расходы, связанные с модулем кросс-модального внимания, компенсируются применением сериализации с комбинацией кривых Пеано в виде кривых Гильберта, Мортона и их транспонированных вариантов.

Заключение

Предложенный в настоящей работе метод обучения мультимодальных моделей компьютерного зрения показал преимущество применения кривых Пеано для сериализации признаков мультимодальных данных. Использование комбинаций кривых Мортона, Гильберта позволило производить кросс-модальное внимание со смешанными признаками облаков точек и изображений быстрее по сравнению с альтернативными подходами. Проведенный анализ возможности ранней сериализации изображений в кривые Пеано показал избыточные затраты на обработку без особых преимуществ в точности. Данный результат объясняется плотной структурой изображений, для которых преобразование координат из плоскости в одномерную кривую не снижает вычислительные затраты. По результатам эксперимента было принято решение производить позднюю сериализацию, на этапе смешения данных с признаками облаков точек. В будущей работе планируется адаптировать представление изображений в более эффективном формате для ранней сериализации, что позволит унифицировать процесс извлечения признаков и провести тестирование на бóльших, по сравнению с CIFAR-10, датасетах.

Эксперименты показали, что VFM DINOv3 может служить учителем в задаче дистилляции знаний для ускорения процесса обучения модели. При этом коэффициент Отиаи является основной функцией потерь для дистилляции знаний из локальных карт признаков DINOv3 в энкодер и декодер Point Transformer v3. Поскольку DINOv3 используется исключительно на стадии предобучения, в режиме исполнения, VFM не вносит никаких дополнительных вычислительных задержек.

Финансовая поддержка

Исследование выполнено в рамках бюджетной темы FFZF-2025-0003.

Литература

1. **Shen W., Peng Z., Wang X., Wang H., Cen J., Jiang D., Xie L., Yang X., Tian Q.** A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 8, pp. 9284–9305. doi:10.1109/TPAMI.2023.3246102
2. **Zhu X., Goldberg A. B.** *Introduction to semi-supervised learning*. Ser.: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3. Morgan & Claypool Publishers, 2009, pp. 1–130. doi:10.2200/S00196ED1V01Y200906AIM006
3. **Zhang Y., Hou J.** Fine-grained Image-to-LiDAR contrastive distillation with Visual Foundation Models. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. doi:10.48550/arXiv.2405.14271
4. **Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., Girshick R.** Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
5. **Siméoni O., Vo H. V., Seitzer M., Baldassarre F., Oquab M., Jose C., Khalidov V., Szafraniec M., Yi S., Ramamonjisoa M., Massa F., Haziza D., Wehrstedt L., Wang J., Darcet T., Moutakanni T., Sentana L., Roberts C., Vedaldi A., Tolan J., Brandt J., Couprie C., Mairal J., Jégou H., Labatut P., Bojanowski P.** DINOv3. 2025. arXiv:2508.10104. doi:10.48550/arXiv.2508.10104
6. **Kamath A., Ferret J., Pathak S., Vieillard N., Merhej R., Perrin S., Matejovicova T., Ramé A., Riviére M., Rouillard L., Mesnard T., Cideron G., Grill J.-B., Ramos S., Yvinec E., Casbon M., Pot E., Penchev I., Liu G., Visin F., Kenealy K., Beyer L., Zhai X., Tsitsulin A., Busa-Fekete R., Feng A., Sachdeva N., Coleman B., Gao Y., Mustafa B., Barr I., Parisotto E., Tian D., Eyal M., Cherry C., Peter J.-T., Sinopalnikov D., Bhupatiraju S., Agarwal R., Kazemi M., Malkin D., Kumar R., Vilar D., Brusilovsky I., Luo J., Steiner A.** Gemma 3 Technical Report. 2025. arXiv:2503.19786. doi:10.48550/arXiv.2503.19786
7. **Татарникова Т. М., Мокрецов Н. С.** Метод дистилляции знаний для языковых моделей на основе выборочного вмешательства в обучение. *Кибернетика и системный анализ*, 2025, т. 150, № 2, с. 361–365. doi:10.15827/0236-235X.150.361-365
8. **Wang X., Zhang X., Cao Y., Wang W., Shen C., Huang T.** SegGPT: Towards segmenting everything in context. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140. doi:10.1109/ICCV51070.2023.00110
9. **Zeid K. A., Yilmaz K., de Geus D., Hermans A., Adrian D., Linder T., Leibe B.** DINO in the Room: Leveraging 2D foundation models for 3D segmentation. 2025. arXiv:2503.18944. doi:10.48550/arXiv.2503.18944
10. **Yang M., Qi Y., Xiong B., Zhang Z., Liu Y.** Modal mimicking knowledge distillation for monocular three-dimensional object detection. *Engineering Applications of Artificial Intelligence*, 2025, vol. 160, Article 111821. doi:10.1016/j.engappai.2025.111821
11. **Bang G., Choi K., Kim J., Kum D., Choi J. W.** RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15491–15500. doi:10.1109/CVPR52733.2024.01467
12. **Wang H., Bao Y., Pan P., Li Z., Liu X., Yang R., Huang D.** Multi-modal relation distillation for unified 3D representation learning. *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Cham, Springer Nature Switzerland, 2025, pp. 364–381. doi:10.1007/978-3-031-73414-4_21
13. **Huang T., Dong B., Yang Y., Huang X., Lau Rynson W. H., Ouyang W., Zuo W.** CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22100–22110. doi:10.1109/ICCV51070.2023.02025
14. **Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I.** Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*; *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8748–8763. https://proceedings.mlr.press/v139/radford21a.html (дата обращения: 11.06.2025).
15. **Wu K., Peng H., Zhou Z., Xiao B., Liu M., Yuan L., Xuan H., Valenzuela M., Chen X., Wang X., Chao H., Hu H.** TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21970–21980. doi:10.1109/ICCV51070.2023.02008
16. **Xu R., Xiang Z., Zhang C., Zhong H., Zhao X., Dang R., Xu P., Pu T., Liu E.** SCKD: Semi-supervised cross-modality knowledge distillation for 4D radar object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 9, pp. 8933–8941. doi:10.1609/aaai.v39i9.32966
17. **Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I.** Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning*; *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html (дата обращения: 11.06.2025).
18. **Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B.** High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695. doi:10.1109/cvpr52688. 2022.01042

19. **Xiao B., Wu H., Xu W., Dai X., Hu H., Lu Y., Zeng M., Liu C., Yuan L.** Florence-2: Advancing a unified representation for a variety of vision tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, June 2024, pp. 4818–4829. doi:10.1109/CVPR52733.2024.00461
20. **Wang W., Lv Q., Yu W., Hong W., Qi J., Wang Y., Ji J., Yang Z., Zhao L., Song X., Xu J., Xu B., Li J., Dong Y., Ding M., Tang J.** CogVLM: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Curran Associates, Inc., 2024, pp. 121475–121499. doi:10.52202/079017-3860
21. **Kim M. J., Pertsch K., Karamcheti S., Xiao T., Balakrishna A., Nair S., Rafailov R., Foster E. P., Sanketi P. R., Vuong Q., Kollar T., Burchfiel B., Tedrake R., Sadigh D., Levine S., Liang P., Finn C.** OpenVLA: An open-source vision-language-action model. *Proceedings of the 8th Conference on Robot Learning*; *Proceedings of Machine Learning Research (PMLR)*, P. Agrawal, O. Kroemer, W. Burgard (Eds.), November 2025, vol. 270, pp. 2679–2713. https://proceedings.mlr.press/v270/kim25c.html (дата обращения: 11.03.2025).
22. **Bai S., Chen K., Liu X., Wang J., Ge W., Song S., Dang K., Wang P., Wang S., Tang J., Zhong H., Zhu Y., Yang M., Li Z., Wan J., Wang P., Ding W., Fu Z., Xu Y., Ye J., Zhang X., Xie T., Cheng Z., Zhang H., Yang Z., Xu H., Lin J.** Qwen2.5-VL Technical Report. 2025. arXiv:2502.13923. doi:10.48550/arXiv.2502.13923
23. **Wu X., Jiang L., Wang P.-S., Liu Z., Liu X., Qiao Y., Ouyang W., He T., Zhao H.** Point Transformer V3: Simpler, faster, stronger. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4840–4851. doi:10.1109/cvpr52733.2024.00463
24. **Бережнов Н. И., Сирота А. А.** Совершенствование механизмов внимания для архитектуры трансформер в задачах повышения качества изображений. *Компьютерная оптика*, 2024, т. 48, № 5, с. 726–733. doi:10.18287/2412-6179-CO-1393
25. **Kutscher D., Chan D. M., Bai Y., Darrell T., Gupta R.** REOrdering patches improves vision models. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. https://openreview.net/forum?id=g56WiaXKGF (дата обращения: 12.01.2026).
26. **Иванов В. Ф., Охотников А. Л., Градусов А. Н.** Алгоритм комплексирования сенсорных данных для задач автоматического управления подвижным составом. *Автоматика на транспорте*, 2024, т. 10, № 4, с. 360–371. doi:10.20295/2412-9186-2024-10-04-360-371, EDN: QWNIRH
27. **Yang Y., Gao X., Wang T., Hao X., Shi Y., Tan X., Ye X., Wang J.** Explore the LiDAR-camera dynamic adjustment fusion for 3D object detection. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 6291–6298. doi:10.48550/arxiv.2407.15334
28. **Daou S., Ben-Hamadou A., Rekik A., Kallel A.** Cross-attention fusion of visual and geometric features for large-vocabulary Arabic lipreading. *Technologies*, 2025, vol. 13, no. 1. doi:10.3390/technologies13010026
29. **Shi H., Wang X., Zhao J., Hua X.** A cross-modal attention-driven multi-sensor fusion method for semantic segmentation of point clouds. *Sensors*, 2025, vol. 25, no. 8. doi:10.3390/s25082474
30. **Krizhevsky A., Nair V., Hinton G.** *CIFAR-10 and CIFAR-100 datasets.* https://www.cs.toronto.edu/~kriz/cifar.html (дата обращения: 9.12.2025).
31. **Simonyan K., Zisserman A.** Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. arXiv:1409.1556v6. https://doi.org/1048550/arXiv:1409.1556
32. **He K., Zhang X., Ren S., Sun J.** Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90
33. **Peng Z., Zheng Y., Cheng Y., Li Y.** Transformer-based LiDAR-camera fusion for semantic segmentation. *2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 2025, pp. 241–245. doi:10.1109/ICBASE66587.2025.11181322
34. **Zhou Z., Ye D., Chen W., Xie Y., Wang Y., Panqu Wang, Foroosh H.** LiDARFormer: A unified transformer-based multi-task network for LiDAR perception. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14740–14747. doi:10.1109/ICRA57147.2024.10610374
35. **Liang T., Xie H., Yu K., Xia Z., Lin Z., Wang Y., Wang B., Tang Z.** BEVFusion: A simple and robust LiDAR-camera fusion framework. *Neural Information Processing Systems (NeurIPS)*, 2022. doi:10.13140/rg.2.2.31757.20966/1

UDC 004.8
doi:10.31799/1684-8853-2026-3-35-48
EDN: XBAJEX

Method for training computer vision models based on cross-modal knowledge distillation using large visual models

A. V. Kuchkov[a], Post-Graduate Student, orcid.org/0009-0007-7508-3348
A. M. Kashevnik[a,b], PhD, Tech., Associate Professor, orcid.org/0000-0001-6503-1447, alexey.kashevnik@iias.spb.su
[a]ITMO University, 49, Kronverksky Pr., 197101, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg Institute for Informatics and Automation of the RAS, 39, 14 Line, V. O., 199178, Saint-Petersburg, Russian Federation

Introduction: Modern methods for training multimodal computer vision models mostly utilize separate feature extraction branches with late fusion. This approach is well suited for integrating existing networks into multimodal systems; however, it is resource-intensive at runtime due to the duplication of processing branches. **Purpose:** To develop a method for building multimodal computer vision models that employs a unified representation of multimodal data in order to simplify data fusion and knowledge distillation. **Methods:** Serialization of sparse and dense data types; cross-modal knowledge distillation for computer vision architectures; utilization of visual foundation models for knowledge distillation in a serialized feature space. **Results:** We develop a method for training computer vision models based on Peano curves using knowledge distillation from large visual models. The method enables blending data of different dimensions using cross-modal attention in real time by applying one-dimensional Peano curves (Gilbert and Morton curves) to serialize multidimensional data. The proposed method demonstrates a latency of 50 ms compared to 40 ms in the single-modal mode (Point Transformer v3), indicating low overhead when using cross-modal distillation on serialized feature maps. The method has been tested in pretraining mode on the nuScenes dataset using the large DINOv3 visual model. In distillation mode, using 25% of the total dataset has yielded 79.2 mIoU compared to 82.1 mIoU on 100% of the dataset using the cosine similarity distillation error function. **Practical relevance:** The use of a serialized data representation allows for accelerated computations on sparse spatial data and makes cross-modal data blending methods less resource-intensive. **Discussion:** The proposed method enables the implementation of cross-modal attention without significant additional computational costs. However, experiments conducted to date have shown that using a serialized encoder for images is impractical due to the inherently dense structure of images. Implementing an image serialization method with a faster execution time would eliminate the need for separate encoders for point cloud and image branches, significantly simplifying the architecture..

Keywords — multimodal models, knowledge distillation, visual foundation models, Peano curves, cross-modal attention.

For citation: Kuchkov A. V., Kashevnik A. M. Method for training computer vision models based on cross-modal knowledge distillation using large visual models. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2026, no. 3, pp. 35–48 (In Russian). doi:10.31799/1684-8853-2026-3-35-48, EDN: XBAJEX

Financial support
This work was funded by the Russian State Research FFZF-2025-0003.

References

1. Shen W., Peng Z., Wang X., Wang H., Cen J., Jiang D., Xie L., Yang X., Tian Q. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no. 8, pp. 9284–9305. doi:10.1109/TPAMI.2023.3246102
2. Zhu X., Goldberg A. B. *Introduction to semi-supervised learning*. Ser.: Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3. Morgan & Claypool Publishers, 2009, pp. 1–130. doi:10.2200/S00196ED1V01Y200906AIM006
3. Zhang Y., Hou J. Fine-grained Image-to-LiDAR contrastive distillation with Visual Foundation Models. *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. doi:10.48550/arXiv.2405.14271
4. Kirillov A., Mintun E., Ravi N., Mao H., Rolland C., Gustafson L., Xiao T., Whitehead S., Berg A. C., Lo W.-Y., Dollár P., Girshick R. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.
5. Siméoni O., Vo H. V., Seitzer M., Baldassarre F., Oquab M., Jose C., Khalidov V., Szafraniec M., Yi S., Ramamonjisoa M., Massa F., Haziza D., Wehrstedt L., Wang J., Darcet T., Moutakanni T., Sentana L., Roberts C., Vedaldi A., Tolan J., Brandt J., Couprie C., Mairal J., Jégou H., Labatut P., Bojanowski P. DINOv3. 2025. arXiv:2508.10104. doi:10.48550/arXiv.2508.10104
6. Kamath A., Ferret J., Pathak S., Vieillard N., Merhej R., Perrin S., Matejovicova T., Ramé A., Riviére M., Rouillard L., Mesnard T., Cideron G., Grill J.-B., Ramos S., Yvinec E., Casbon M., Pot E., Penchev I., Liu G., Visin F., Kenealy K., Beyer L., Zhai X., Tsitsulin A., Busa-Fekete R., Feng A., Sachdeva N., Coleman B., Gao Y., Mustafa B., Barr I., Parisotto E., Tian D., Eyal M., Cherry C., Peter J.-T., Sinopalnikov D., Bhupatiraju S., Agarwal R., Kazemi M., Malkin D., Kumar R., Vilar D., Brusilovsky I., Luo J., Steiner A. Gemma 3 Technical Report. 2025. arXiv:2503.19786. doi:10.48550/arXiv.2503.19786
7. Tatarnikova T. M., Mokretsov N. S. Knowledge distillation method for language models based on selective intervention in training. *Cybernetics and Systems Analysis*, 2025, vol. 150, no. 2, pp. 361–365 (In Russian). doi:10.15827/0236-235X.150.361-365
8. Wang X., Zhang X., Cao Y., Wang W., Shen C., Huang T. SegGPT: Towards segmenting everything in context. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1130–1140. doi:10.1109/ICCV51070.2023.00110
9. Zeid K. A., Yilmaz K., de Geus D., Hermans A., Adrian D., Linder T., Leibe B. DINO in the Room: Leveraging 2D foundation models for 3D segmentation. 2025. arXiv:2503.18944. doi:10.48550/arXiv.2503.18944
10. Yang M., Qi Y., Xiong B., Zhang Z., Liu Y. Modal mimicking knowledge distillation for monocular three-dimensional object detection. *Engineering Applications of Artificial Intelligence*, 2025, vol. 160, Article 111821. doi:10.1016/j.engappai.2025.111821
11. Bang G., Choi K., Kim J., Kum D., Choi J. W. RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features. *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15491–15500. doi:10.1109/CVPR52733.2024.01467
12. Wang H., Bao Y., Pan P., Li Z., Liu X., Yang R., Huang D. Multi-modal relation distillation for unified 3D representation learning. *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, G. Varol (Eds.), Cham, Springer Nature Switzerland, 2025, pp. 364–381. doi:10.1007/978-3-031-73414-4_21
13. Huang T., Dong B., Yang Y., Huang X., Lau Rynson W. H., Ouyang W., Zuo W. CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training. *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22100–22110. doi:10.1109/ICCV51070.2023.02025
14. Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G., Sutskever I. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*; *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8748–8763. Available at: https://proceedings.mlr.press/v139/radford21a.html http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part4/op (accessed 11 June 2025).
15. Wu K., Peng H., Zhou Z., Xiao B., Liu M., Yuan L., Xuan H., Valenzuela M., Chen X., Wang X., Chao H., Hu H. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 21970–21980. doi:10.1109/ICCV51070.2023.02008
16. Xu R., Xiang Z., Zhang C., Zhong H., Zhao X., Dang R., Xu P., Pu T., Liu E. SCKD: Semi-supervised cross-modality knowledge distillation for 4D radar object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 9, pp. 8933–8941. doi:10.1609/aaai.v39i9.32966
17. Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I. Zero-shot text-to-image generation. *Proceedings of the 38th International Conference on Machine Learning*. In: *Proceedings of Machine Learning Research (PMLR)*, M. Meila, T. Zhang (Eds.), Jule 2021, vol. 139, pp. 8821–8831. Available at: https://proceedings.mlr.press/v139/ramesh21a.html (accessed 11 June 2025).

18. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695. doi:10.1109/cvpr52688.2022.01042
19. Xiao B., Wu H., Xu W., Dai X., Hu H., Lu Y., Zeng M., Liu C., Yuan L. Florence-2: Advancing a unified representation for a variety of vision tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, IEEE Computer Society, June 2024, pp. 4818–4829. doi:10.1109/CVPR52733.2024.00461
20. Wang W., Lv Q., Yu W., Hong W., Qi J., Wang Y., Ji J., Yang Z., Zhao L., Song X., Xu J., Xu B., Li J., Dong Y., Ding M., Tang J. CogVLM: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Curran Associates, Inc., 2024, pp. 121475–121499. doi:10.52202/079017-3860
21. Kim M. J., Pertsch K., Karamcheti S., Xiao T., Balakrishna A., Nair S., Rafailov R., Foster E. P., Sanketi P. R., Vuong Q., Kollar T., Burchfiel B., Tedrake R., Sadigh D., Levine S., Liang P., Finn C. OpenVLA: An open-source vision-language-action model. *Proceedings of the 8th Conference on Robot Learning*. In: *Proceedings of Machine Learning Research (PMLR)*, P. Agrawal, O. Kroemer, W. Burgard (Eds.), November 2025, vol. 270, pp. 2679–2713. Available at: https://proceedings.mlr.press/v270/kim25c.html (accessed 11 March 2025).
22. Bai S., Chen K., Liu X., Wang J., Ge W., Song S., Dang K., Wang P., Wang S., Tang J., Zhong H., Zhu Y., Yang M., Li Z., Wan J., Wang P., Ding W., Fu Z., Xu Y., Ye J., Zhang X., Xie T., Cheng Z., Zhang H., Yang Z., Xu H., Lin J. Qwen2.5-VL Technical Report. 2025. arXiv:2502.13923. doi:10.48550/arXiv.2502.13923
23. Wu X., Jiang L., Wang P.-S., Liu Z., Liu X., Qiao Y., Ouyang W., He T., Zhao H. Point Transformer V3: Simpler, faster, stronger. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4840–4851. doi:10.1109/cvpr52733.2024.00463
24. Berezhnov N. I., Sirota A. A. Improving attention mechanisms in transformer architecture in image restoration. *Computer Optics*, 2024, vol. 48, no. 5, pp. 726–733 (In Russian). doi:10.18287/2412-6179-CO-1393
25. Kutscher D., Chan D. M., Bai Y., Darrell T., Gupta R. REOrdering patches improves vision models. *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. Available at: https://openreview.net/forum?id=g-56WiaXKGF (accessed 12 January 2026).
26. Ivanov V. F., Ohotnikov A. L., Gradusov A. N. Algorithm of complexing sensor data for tasks of automatic control of rolling stock. *Transport Automation Research*, 2024, vol. 10, no. 4, pp. 360–371 (In Russian). doi:10.20295/2412-9186-2024-10-04-360-371, EDN: QWNIRH
27. Yang Y., Gao X., Wang T., Hao X., Shi Y., Tan X., Ye X., Wang J. Explore the LiDAR-camera dynamic adjustment fusion for 3D object detection. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2025, pp. 6291–6298. doi:10.48550/arxiv.2407.15334
28. Daou S., Ben-Hamadou A., Rekik A., Kallel A. Cross-attention fusion of visual and geometric features for large-vocabulary Arabic lipreading. *Technologies*, 2025, vol. 13, no. 1. doi:10.3390/technologies13010026
29. Shi H., Wang X., Zhao J., Hua X. A cross-modal attention-driven multi-sensor fusion method for semantic segmentation of point clouds. *Sensors*, 2025, vol. 25, no. 8. doi:10.3390/s25082474
30. Krizhevsky A., Nair V., Hinton G. *CIFAR-10 and CIFAR-100 datasets*. Available at: https://www.cs.toronto.edu/~kriz/cifar.html (accessed 9 December 2025).
31. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. arXiv: 1409.1556v6. https://doi.org/1048550/ arXiv:1409.1556
32. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90
33. Peng Z., Zheng Y., Cheng Y., Li Y. Transformer-based LiDAR-camera fusion for semantic segmentation. *2025 6th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 2025, pp. 241–245. doi:10.1109/ICBASE66587.2025.11181322
34. Zhou Z., Ye D., Chen W., Xie Y., Wang Y., Panqu Wang, Foroosh H. LiDARFormer: A unified transformer-based multi-task network for LiDAR perception. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14740–14747. doi:10.1109/ICRA57147.2024.10610374
35. Liang T., Xie H., Yu K., Xia Z., Lin Z., Wang Y., Wang B., Tang Z. BEVFusion: A simple and robust LiDAR-camera fusion framework. *Neural Information Processing Systems (NeurIPS)*, 2022. doi:10.13140/rg.2.2.31757.20966/1