

Огляд багатосенсорного сприйняття на основі злиття даних для втіленого штучного інтелекту: передумови, методи, виклики та перспективи

Shulan Ruan, Rongwei Wang, Xuchen Shen, Huijie Liu, Baihui Xiao, Jun Shi, Kun Zhang, Zhenya Huang, Yu Liu, Enhong Chen, You He

Анотація—Багатосенсорне сприйняття на основі злиття даних (multi-sensor fusion perception, MSFP) є ключовою технологією для втіленого штучного інтелекту (ШІ), яка обслуговує різноманітні низхідні завдання (напр., 3D-детектування об'єктів і семантичну сегментацію) та сценарії застосування (напр., автономне керування та роботи у складі рою). Останнім часом значні досягнення методів MSFP на основі ШІ було розглянуто у відповідних оглядових працях. Однак, як ми спостерігаємо після ретельного й детального дослідження, наявні огляди мають низку обмежень. По-перше, більшість оглядів орієнтовано на окреме завдання або окрему дослідницьку область, наприклад, 3D-детектування об'єктів або автономне керування. Тому дослідникам у суміжних задачах часто складно отримати з них пряму користь. По-друге, більшість оглядів подають MSFP лише з однієї перспективи мультимодального злиття, не враховуючи різноманітність методів MSFP, як-от багатовидове злиття або злиття часових рядів. Зважаючи на це, у даній роботі ми прагнемо впорядкувати дослідження MSFP з незалежної від завдання перспективи, де методи представлено з різних технічних точок зору. Зокрема, спершу ми вводимо передумови MSFP. Далі розглядаємо методи мультимодального та мультиагентного злиття. Далі аналізуємо методи злиття часових рядів. В епоху великих мовних моделей (LLM) ми також досліджуємо мультимодальні методи злиття на основі LLM (MM-LLM). Нарешті, обговорюємо відкриті виклики та перспективні напрямки для MSFP. Сподіваємось, що цей огляд допоможе дослідникам зрозуміти важливий поступ у MSFP і дасть можливі відправні точки для майбутніх досліджень.

Index Terms—багатосенсорне сприйняття на основі злиття даних, втілений ШІ, мультимодальність, мультимодальний, часові ряди, MM-LLM

I. Вступ

Shulan Ruan, Rongwei Wang, Xuchen Shen, Baihui Xiao, Yu Liu, and You He are with Tsinghua University (e-mail: slruan@sz.tsinghua.edu.cn)

Huijie Liu, Jun Shi, Zhenya Huang, and Enhong Chen are with University of Science and Technology of China
Kun Zhang is with Hefei University of Technology

А останні роки, завдяки стрімкому розвитку глибокого навчання та великих мовних моделей (large language model, LLM), штучний інтелект (ШІ) досяг помітного прогресу в різноманітних галузях [?], [?], [?]. Як важливий напрямок ШІ, *втілений ШІ* (embodied AI) — це інтелект, що використовує фізичні сутності як носії й реалізує здатність до автономного прийняття рішень і дій через сприйняття в реальному часі в динамічному середовищі. Втілений ШІ має широкий спектр сценаріїв застосування, як-от автономне керування або колективна інтелектуальна поведінка роботів-роїв [?], [?]. В останні роки він став важливою темою досліджень у спільноті ШІ та одним із ключових шляхів подолання поточних обмежень розвитку ШІ й досягнення загального штучного інтелекту (AGI).

У побудові систем втіленого ШІ розуміння сенсорних даних — це центральна ланка між фізичним світом і цифровим інтелектом. На відміну від традиційного візуально-домінованого режиму сприйняття, втілені агенти повинні інтегрувати мультимодальні сенсорні дані задля панорамного сприйняття середовища. Серед таких сенсорів — візуальні камери, радары міліметрового діапазону, LiDAR, інфрачервоні камери та інерціальні вимірювальні модулі (IMU). Багатосенсорне сприйняття на основі злиття даних (MSFP) є вирішальним для досягнення стійкого сприйняття і точного прийняття рішень у втіленому ШІ. Наприклад, візуальні камери легко піддаються впливу змін освітлення, а характеристики LiDAR суттєво погіршуються в умовах дощу й туману.

Як показано на рис. ??, поточні дослідження багатосенсорного сприйняття на основі злиття даних для втіленого інтелекту здебільшого ґрунтуються на парадигмі “Агент-Сенсор-Дані-Модель-Завдання”.

Існуючі методи MSFP вже досягли вражаючих результатів у багатьох областях, як-от автономне керування та промислова робототехніка, але їх застосування до втіленого ШІ все ще стикається з низкою принципових викликів. Зокрема, по-перше, гетерогенність

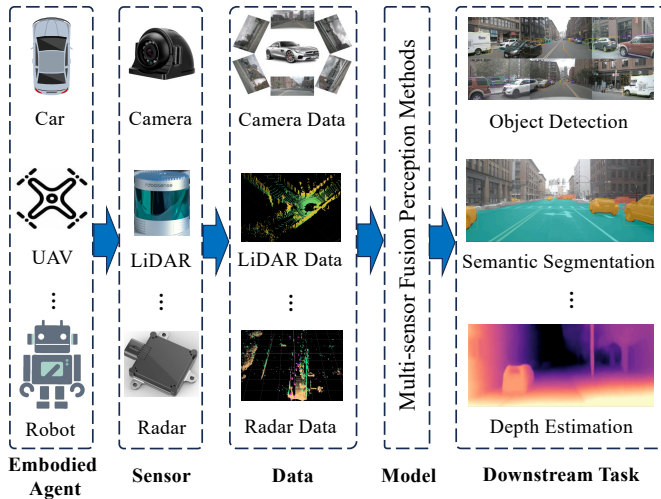


Рис. 1. Загальна схема конвеєра багатосенсорного сприйняття на основі злиття даних.

крос-модальних даних ускладнює уніфікацію простору ознак. По-друге, просторово-часова асинхронність між різними сенсорами може спричинити помилки злиття. Крім того, відмова сенсора (напр., забруднення оптики або перекриття сигналу) може спричинити динамічну втрату мультимодальної інформації.

Навколо цих проблем, як показано у табл. ??, за останні роки з'явилися огляди, що систематично узагальнюють відповідні методи [?], [?], [?], [?], [?], [?], [?], [?]. Попри значні зусилля, наше детальне дослідження показує, що ці огляди мають низку обмежень. По-перше, більшість оглядів зорієнтовано на одне завдання або одну дослідницьку область, наприклад, 3D-детектування об'єктів чи автономне керування. Через це дослідники у суміжних задачах часто не можуть скористатися ними напряму. По-друге, більшість оглядів подають MSFP лише з однієї перспективи мультимодального злиття, не враховуючи різноманітності методів MSFP, як-от мультиагентне злиття та злиття часових рядів.

Тож у цій статті ми прагнемо впорядкувати дослідження MSFP з незалежної від завдання перспективи, де методи представлено суто з різних технічних поглядів. Зокрема, ми спочатку вводимо передумови MSFP, включно з різними завданнями сприйняття, різними типами сенсорних даних, поширеними наборами даних та відповідними критеріями оцінювання. Далі ми розглядаємо методи мультимодального злиття з рівнів окремих точок, вокселів, регіонів та багаторівневого злиття. Послідовно ми досліджуємо методи мультиагентного злиття, зосереджуючись на спільному (коллаборативному) сприйнятті між кількома втіленими агентами та інфраструктурою. Далі також досліджуються методи злиття часових рядів,

Табл. I
Огляд споріднених оглядів з MSFP.

Огляд	Рік	Область	Завдання
Wang <i>et al.</i> [?]	2023	Автономне керування	3D-детектування об'єктів
Wang <i>et al.</i> [?]	2023	Автономне керування	3D-детектування об'єктів
Xiang <i>et al.</i> [?]	2023	Автономне керування	Не специфіковано
Zhu <i>et al.</i> [?]	2023	Не специфіковано	SLAM
Tang <i>et al.</i> [?]	2023	Автономне керування	3D-детектування об'єктів
Du <i>et al.</i> [?]	2024	Не специфіковано	SLAM
Bin <i>et al.</i> [?]	2024	Гуманоїдні роботи	Не специфіковано
Song <i>et al.</i> [?]	2024	Автономне керування	3D-детектування об'єктів
Han <i>et al.</i> [?]	2025	Робототехніка	Не специфіковано

дів, які об'єднують часово-впорядковані (напр., кілька попередніх кадрів) сенсорні дані для прогнозування. В епоху великих моделей ми досліджуємо методи злиття на основі MM-LLM, поділяючи їх на "візуально-мовні" (vision-language) та "візуально-LiDAR-мовні" (vision-LiDAR-language); ці методи рідко включалися до попередніх оглядів. Нарешті, ми всебічно обговорюємо відкриті виклики та майбутні можливості у MSFP на рівнях даних, моделі та застосування. Сподіваємось, цей огляд допоможе дослідникам зрозуміти важливий поступ у MSFP за останнє десятиліття та надасть відповідні точки для майбутніх досліджень.

Решту статті організовано так. У розділі ?? ми описуємо передумови MSFP з боку різних сенсорних даних, доступних наборів даних і різних завдань сприйняття. У розділі ?? ми вводимо методи мультимодального злиття на різних рівнях, напр., рівень окремих точок, вокселів, регіонів і багаторівневий. У розділі ?? ми підсумовуємо методи мультиагентного спільного сприйняття. У розділі ?? ми розглядаємо методи злиття часових рядів для MSFP. У розділі ?? ми досліджуємо сучасні методи MSFP на основі MM-LLM. У розділі ?? ми обговорюємо відкриті виклики та майбутні напрямки в MSFP. Нарешті, ми завершуємо нашу роботу в розділі ??.

II. Передумови

У цьому розділі ми зосереджуємось на передумовах MSFP. Спершу ми досліджуємо поширені сенсори та їхні типи даних. Далі систематизуємо доступні дослідникам еталонні набори даних. Нарешті, докладно описуємо різноманітні низхідні завдання для MSFP.

Табл. II

Статистика популярних наборів даних для MSFP. Тут L, R, C означають LiDAR, Radar та Camera (камеру) відповідно. U, S, H означають міські, приміські та шосейні умови відповідно.

Набір даних	Модальність	Сценарій	# Класів	# Кадрів
KITTI [?]	L + C	U + S + H	8	15K
nuScenes [?]	L + R + C	U + S	23	15K
Waymo Open [?]	L + R + C	U + S	4	1M
Cityscapes 3D [?]	C	U	30	35K
Argoverse [?]	L + R + C	U	15	1.75M
A*3D [?]	L + C	U	10	240K
ApolloScape [?]	L + C	U + S + H	35	500K+
AIODrive [?]	L + R + C	U	20	60K
H3D [?]	L + C	U	8	

II-A. Сенсорні дані

II-A1. Дані з камер: Камери здатні фіксувати багаті ознаки зовнішнього вигляду об'єктів — кольори, форми та текстури, що є вирішальним для різноманітних завдань сприйняття. Однак, як пасивні сенсори, камери чутливі до умов освітлення. Якість зображення суттєво погіршується вночі та за несприятливої погоди, як от туман і дощ.

II-A2. Дані LiDAR: LiDAR обчислює відстані до об'єктів, вимірюючи часову різницю між випроміненим і прийнятим лазерним сигналом. Він безпосередньо видає високоточні 3D-хмари точок, що містять просторову геометричну інформацію, що дає унікальні переваги в 3D-сприйнятті. Втім, він зазвичай чутливий до погоди. Через властиві розрідженість і нерівномірність, ефективне представлення та розуміння хмар точок LiDAR також залишаються складними задачами.

II-A3. Дані радарів міліметрового діапазону: Радари міліметрового діапазону (mmWave) детектують об'єкти, випромінюючи й приймаючи радіохвилі. Порівняно з хмарами точок LiDAR, хмари точок радарів є розрідженішими й погано описують контури об'єктів. Однак радари зберігають хорошу продуктивність у несприятливих погодних умовах і можуть безпосередньо вимірювати швидкості об'єктів.

II-B. Набори даних

II-B1. KITTI: KITTI [?] складається з 14 999 зображень і відповідних хмар точок, з яких 7 481 для навчання і 7 518 для тестування. Анотації охоплюють вісім категорій та поділяються на прості, середні та складні залежно від розміру, перекриттів і рівнів обрізання. Транспортний засіб для збору даних мав дві монохромні камери, дві кольорові камери, 64-променевий LiDAR Velodyne, чотири оптичні об'єктиви та GPS-систему. Дані зібрано приблизно з 50 сцен у Карлсруе та сусідніх німецьких містах, що охоплюють міські, сільські й шосейні умови.

II-B2. nuScenes: NuScenes [?] було зібрано в Бостоні та Сінгапурі. Містить 700 тренувальних, 150 валідаційних і 150 тестових сцен. Кожна сцена триває приблизно 20 секунд із 40 зразками, загалом близько 5,5 годин. Набір містить 1,4 мільйона зображень з камер, 390 тис. сканувань LiDAR, 1,4 мільйона сканувань радарів та 1,4 мільйона анотованих обмежувальних рамок у 40 тис. ключових кадрах. Платформа основана на 6 камерах з оглядом 360 градусів, 324-променевим LiDAR з 1,39 мільйона точок на кадрі, 5 радарів міліметрового діапазону та інерціальною навігаційною системою з GPS і IMU.

II-B3. Waymo Open: Waymo Open [?] містить набори для сприйняття та руху. Анотації у наборі для сприйняття включають 1,26 мільйона 3D-обмежувальних рамок, 1,18 мільйона 2D-обмежувальних рамок, мітки паноптичної сегментації для 100 тис. зображень, 14 ключових точок (keypoints) та мітки 3D-семантичної сегментації. Набір для руху містить 103 354 кліпи з траєкторіями об'єктів. Набір охоплює сценарії денного, нічного, світанкового, сутінкового та дощового часу, проте бракує екземплярів екстремальної погоди.

II-B4. Cityscapes 3D: Cityscapes 3D [?] походить від набору Cityscapes [?], доповненого анотаціями 3D-обмежувальних рамок. Він складається з 5 тис. зображень із тонкими анотаціями (тобто 2048×1024 пікселів) і 20 тис. зображень з грубими анотаціями, що використовуються для завдань 3D-розуміння сцени міських ландшафтів, напр., семантичної сегментації на рівні екземпляра.

II-B5. Argoverse: Argoverse [?] зібрано двома 32-канальними LiDAR-сенсорами, сімома круговими камерами та двома передніми стереокамерами, що охоплюють 360 градусів. Він містить 3D-набір для трекінгу із 1 тис. 3D-анотованих зразків, що охоплює 30 категорій об'єктів. Також містить набір для передбачення руху з 250 тис. зразків з траєкторними даними сцен, 20 тис. неанотованих LiDAR-зразків і 1 тис. карт високої роздільної здатності, що надають багату семантичну анотаційну інформацію щодо дорожньої інфраструктури та правил дорожнього руху.

II-B6. A*3D: A*3D [?] зібрано переважно на міських дорогах Сінгапуру. Містить понад 39 тис. анотованих кадрів, кожен з яких помічений 2D- та 3D-обмежувальними рамками і має ідентифікатор трекінгу між кадрами. Набір A*3D включає різні сенсорні дані — щільні 3D-хмари точок, RGB-зображення високої роздільної здатності та IMU-дані, що покривають огляд 360 градусів. Він охоплює різні погодні умови (день, ніч, дощ) та різні сценарії збору даних на

міських дорогах.

II-B7. ApolloScape: ApolloScape [?] зібрано двома LiDAR-сенсорами, шістьма відеокамерами та системою IMU/GNSS; містить понад 140 тис. зображень високої роздільної здатності. Покриває різні часові інтервали та погодні умови; має загалом 25 категорій.

II-B8. AIODrive: AIODrive [?] розроблено дослідницькою групою Університету Карнегі-Меллона; теж зорієнтовано на міську сцену. Синтетичні сенсорні дані AIODrive включають п'ять RGB-камер 1920×720 і п'ять камер глибини, один радар, один LiDAR Velodyne-64, IMU, GPS та три далекодійні LiDAR високої щільності.

II-B9. H3D: H3D [?] зосереджено переважно на 3D-детектуванні та трекінгу об'єктів у міських середовищах. Надає приблизно 160 міських сцен, зібраних LiDAR Velodyne HDL-64E та трьома RGB-камерами високої роздільної здатності, загалом приблизно 27 тис. кадрів. Кожен кадр містить детальні 3D-обмежувальні рамки та ідентифікатори трекінгу об'єктів.

II-C. Завдання сприйняття

II-C1. Детектування об'єктів: Детектування об'єктів — одне з фундаментальних завдань широких систем сприйняття; його основна мета — точно локалізувати та ідентифікувати різні типи об'єктів за даними сенсорів. У 2D-детектуванні система повинна видавати інформацію про категорію об'єкта та 2D-обмежувальну рамку, представлену як (x, y, w, h) . У сценарії 3D-детектування результати детектування мають включати 3D-координати позиції (x, y, z) , інформацію про 3D-розмір (l, w, h) та кут орієнтації (курс) θ цілі.

II-C2. Семантична сегментація: Завдання семантичної сегментації полягає у тому, щоб класифікувати кожну базову одиницю сцени, наприклад піксель зображення, в одну із семантичних категорій. Зокрема, для вхідного набору даних (напр., множини пікселів зображення $I = \{I_1, I_2, \dots, I_n\}$) і попередньо визначеної множини семантичних категорій $y = \{y_1, y_2, \dots, y_k\}$, модель сегментації повинна призначити кожній базовій одиниці I_i відповідну семантичну мітку або розподіл імовірностей класів.

II-C3. Оцінювання глибини: Оцінювання глибини має на меті отримати інформацію про глибину сцени за сенсорними даними та надати втіленим агентам 3D-геометричне розуміння. За заданим вхідним зображенням $I \in \mathcal{R}^{M \times N}$ і відповідною розрідженою картою глибини $D_s \in \mathcal{R}^{M \times N}$ система оцінювання глибини повинна видати щільну карту глибини D_d , де процес допов-

Табл. III
Класифікація методів мультимодального злиття для MSFP.

Категорія	Особливість	Методи
Злиття на рівні точок	Інтегрують геометричні координати точок LiDAR із семантичними деталями зображень на рівні окремих точок.	PointFusion [?], PointPainting [?], MVP [?], DeepFusion [?]
Злиття на рівні вокселів	Перетворюють нерегулярні хмари точок LiDAR на регулярні воксельні ґратки для ефективної обробки; зберігають геометричну інформацію, водночас інтегруючи мультимодальні дані для підвищення семантичного багатства.	CenterFusion [?], PointAugmenting [?], UVTR [?], SFD [?]
Злиття на рівні регіонів	Агрегують специфічні для регіону ознаки між модальностями через просторове вирівнювання, використовуючи міжмодальні просторові відповідності з пізнім злиттям.	AVOD [?], RoarNet [?], AR-CNN [?], E2E-MFD [?]
Багаторівневе злиття	Поєднують мультимодальне злиття на кількох ієрархічних рівнях, використовуючи такі техніки, як багаторетапне злиття, увага або контрастне навчання, для підвищення стійкості сприйняття.	MVX-Net [?], RCBEV [?], MBNet [?], CSSA [?]

нення глибини можна представити як відображення $D_d = f(I, D_s)$. Через оцінювання глибини система здатна отримати точну 3D-інформацію про положення об'єктів у сцені, що критично важливо для низхідних завдань, як-от планування шляху та керування рішеннями.

II-C4. Передбачення зайнятості: Передбачення зайнятості (occuancy prediction) забезпечує насичене семантичне розуміння 3D-простору. Дискретизуючи неперервний 3D-простір у воксельну ґратку, модель сприйняття зайнятості може передбачити стан зайнятості та семантичну категорію кожного вокселя, надаючи повне представлення сцени для автономних рішень.

III. Методи мультимодального злиття

Зливаючи дані мультимодальних сенсорів, втілені агенти можуть зменшити сліпі плями сприйняття та досягти комплекснішого сприйняття середовища. Наприклад, LiDAR забезпечує точну інформацію про глибину, тоді як камери зберігають більше детальної семантичної інформації. Тому питання, як краще зливати мультимодальні дані з різних сенсорів задля точнішого і стійкішого сприйняття, стало гарячою темою досліджень у численних застосуваннях. Як показано в табл. ??, у цьому

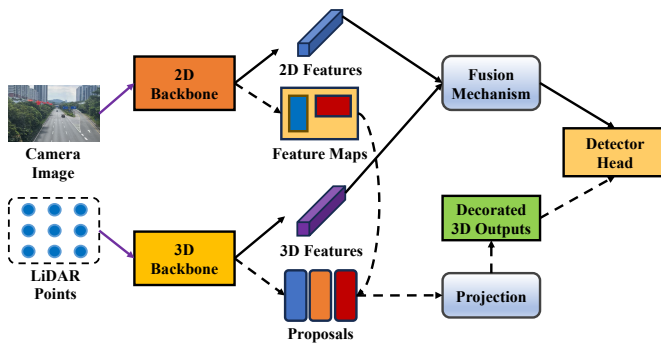


Рис. 2. Загальна схема конвеєра злиття на рівні точок.

розділі ми представляємо різні методи з точки зору різних рівнів злиття, *тобто* рівня точок, вокселів, регіонів та багаторівневого.

III-A. Злиття на рівні точок

Рис. ?? показує типовий конвеєр методів злиття на рівні точок, які мають на меті досягти злиття ознак на рівні окремих точок між хмарами точок LiDAR і даними зображень. Інтегруючи геометричну координатну інформацію хмар точок із семантичними деталями зображень (*напр.*, кольори та категорійні атрибути), можна підвищити точність мультимодального сприйняття. Серед різних методів, PointNet [?] та PointNet++ [?] спочатку обробляли хмари точок безпосередньо, не покладаючись на інші форми представлення, як-от воксели. Вони використовувалися лише для розпізнавання 3D-об'єктів на основі LiDAR. Frustum PointNets [?] розширює PointNet, перетворюючи 2D-кандидатські рамки на 3D-зрізані піраміди (frustums) і виконуючи сегментацію й регресію безпосередньо на сирих хмарах точок. PointFusion [?] застосовує стандартніший підхід, окремо витягаючи ознаки з RGB-зображень і хмар точок за допомогою CNN та PointNet відповідно, а потім конкатенуючи їх для злиття.

Однак це початкове злиття погано вловлює складні крос-модальні зв'язки. PI-RCNN [?] вдосконалює це двостадійним процесом, використовуючи увагу (attentive aggregation) для уточнення злиття 3D-пропозицій і 2D-семантичних ознак, що дає змогу детальнішої обробки. Методи, як-от PointPainting [?] і FusionPainting [?], анують кожну точку LiDAR ознаками зображення: перший проектує точки LiDAR на маски сегментації, другий використовує адаптивну увагу для злиття на семантичному рівні. Ці підходи краще опрацьовують розрідженість хмар точок, ніж методи з пріоритетом пропозицій, як-от PI-RCNN. Аналогічно, MVP [?] збагачує розріджені хмари точок, проектуючи

результати 2D-детектування у віртуальні 3D-точки та зливаючи їх із даними LiDAR, компенсуючи обмеження LiDAR у детектуванні малих або віддалених об'єктів. DeepFusion [?] застосовує механізм перехресної уваги для динамічного вирівнювання ознак LiDAR і зображень і розв'язує проблеми геометричного розузгодження через зворотну аугментацію даних. GraphAlign [?] далі оптимізує процес вирівнювання шляхом графового зіставлення ознак. Він досягає точного зіставлення на рівні пікселів між геометричними ознаками хмари точок і семантичними ознаками зображень через модулі вирівнювання графових ознак та самоуваги, розв'язуючи ключові проблеми неточних геометричних позицій та неоднозначних семантичних асоціацій у мультимодальному злитті.

III-B. Злиття на рівні вокселів

Методи злиття на рівні вокселів перетворюють нерегулярні хмари точок LiDAR на регулярні ґратки (*напр.*, воксели або стовпці-pillars), що дозволяє ефективну обробку зі збереженням геометричної інформації. Рис. ?? показує типовий конвеєр методів злиття на рівні вокселів. Для використання семантичного багатства зображень, зображення камер інтегруються у методи на основі вокселів для покращення сприйняття, особливо у розріджених або зайнятих сценаріях.

Щоб подолати такі проблеми, як неточна інформація про висоту, CenterFusion [?] розширює радарні точки у 3D-стовпці й асоціює радарні детектування з об'єктами на зображеннях. Однак методи на рівні вокселів часто страждають на "розмиття ознак" через втрату просторової інформації всередині вокселя. VPFFNet [?] пом'якшує це, використовуючи шар воксельного RoI-пулінгу та віртуальні точки для вирівнювання та агрегації ознак з LiDAR і зображень. PointAugmenting [?] збагачує точки LiDAR ознаками зображень і вокселізує доповнену хмару точок. Однак проектування 3D-точок на площину зображення може погіршити продуктивність в зайнятих ділянках. VFF [?] вводить метод проектування "точка-в-промінь" (point-to-ray), зливаючи ознаки зображення вздовж променів задля забезпечення багатшого контекстуального інформаційного супроводу — особливо корисно для детектування зайнятих та віддалених об'єктів.

Для вирівнювання ознак, AutoAlign [?] вводить навчальний мультимодальний каркас злиття, динамічно вирівнюючи ознаки зображень і хмар точок без покладання на проекційну матрицю. Як покращена версія AutoAlign,

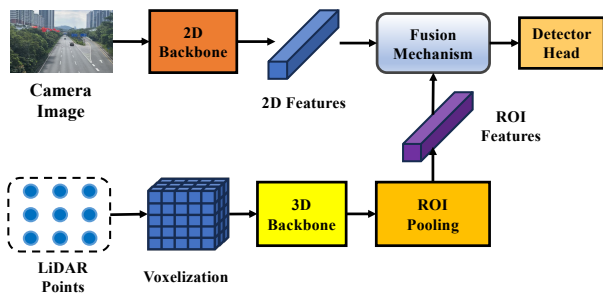


Рис. 3. Загальна схема конвеєра злиття на рівні вокселів.

AutoAlignV2 [?] використовує деформівну увагу та розріджену вибірку для підвищення ефективності та зниження обчислювальних витрат, водночас спрощуючи аугментацію даних порівняно з ранішими методами, як-от PointAugmenting. VoxelNextFusion [?] поєднує воксельні ознаки з відповідними та оточуючими піксельними ознаками за допомогою самоуваги для злиття точок і блоків. Цей підхід ефективно розв'язує невідповідності роздільної здатності та покращує детектування далеких і складних об'єктів.

III-C. Злиття на рівні регіонів

Методи злиття на рівні регіонів зосереджуються на агрегації специфічної для регіонів інформації, такої як карти ознак, ROI (область інтересу) або регіональні пропозиції, з 2D-зображень та інших модальностей. Ці методи особливо ефективні у сценаріях, де просторове вирівнювання між модальностями досягти простіше. AVOD [?] ввів мультимодальну мережу пропозицій регіонів, яка обробляє BEV- та RGB-зображення окремо для генерації карт ознак високої роздільної здатності. Регресуючи вектори напрямку, AVOD розв'язує неоднозначність у визначенні напрямку. Аналогічно, RoarNet [?] застосовує двостадійний каркас: перша стадія передбачає 3D-пози безпосередньо із зображень для уникнення втрати інформації, пов'язаної з проекцією, а друга стадія уточнює ці передбачення з використанням міркувань про хмару точок. TransFusion [?] використовує трансформери для злиття LiDAR-camera, встановлюючи м'які асоціації між точками LiDAR і пікселями зображень. Він адаптується до контекстуальної інформації, розв'язуючи проблеми стійкості, спричинені поганою якістю зображення або помилками калібрування сенсорів.

Для злиття термальних і RGB-зображень методи на рівні регіонів є поширенішими через простіше вирівнювання ознак. Рис. ?? відображає конвеєр методів злиття на рівні регіонів.

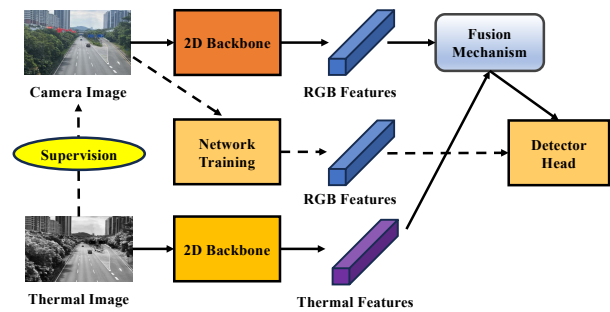


Рис. 4. Загальна схема конвеєра злиття на рівні регіонів.

CMT-CNN [?] вирішує задачу детектування пішоходів за умов слабкого освітлення, реконструюючи термальні регіони, що відповідають RGB-кандидатським регіонам, і зливаючи крос-модальну інформацію через мультимасштабну мережу детектування. AR-CNN [?] розв'язує розузгодження між RGB- та термальними зображеннями, передбачаючи позиційні зсуви та адаптивно вирівнюючи регіональні ознаки. Двопотоккові архітектури, як-от GAFF [?] і RSDet [?], обробляють RGB- і термальні зображення окремо перед злиттям ознак. GAFF використовує внутрішньо- та міжмодальні механізми уваги для вибору ознак, а RSDet уточнює злиття через видалення надмірних спектрів та динамічний вибір ознак.

III-D. Багаторівневе злиття

Багаторівневе злиття інтегрує мультимодальну інформацію з різних рівнів для забезпечення комплекснішого сприйняття. Рис. ?? ілюструє конвеєр методів багаторівневого злиття. У літературі Liang *et al.* [?] використовують неперервні згортки для злиття карт ознак зображень і LiDAR на різних рівнях у просторі BEV. Zhu *et al.* [?] представляють двостадійний метод крос-модального злиття, що збагачує семантичне багатство та локальне представлення пропозицій з рівня точок і рівня регіонів. У такий спосіб продуктивність сприйняття можна покращити у сценаріях з розрідженістю та зайнятістю. Аналогічно, MVX-Net [?] виконує злиття на рівні точок і на рівні вокселів. MMF [?] розширює цю ідею до багатозадачного каркасу, як-от 2D/3D-детектування, оцінювання поверхні та доповнення глибини.

Для підвищення стійкості, EPNet [?] вводить модуль LI-Fusion, що знижує вплив нерелевантної інформації, зливаючи ознаки зображень і хмар точок на різних масштабах. Як покращена версія, EPNet++ [?] також вводить двонапрямну взаємодію інформації: ознаки хмари точок використовуються для уточнення ознак зображень, і навпаки. У такий спо-

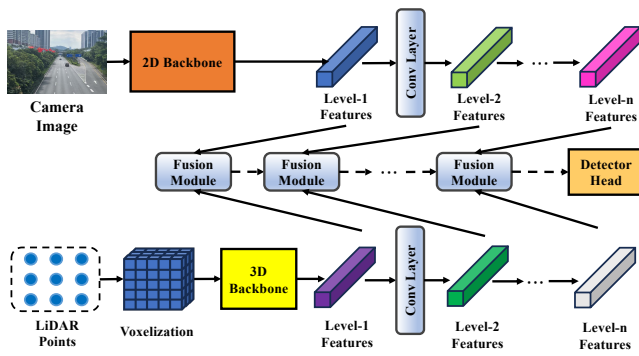


Рис. 5. Конвеєр багаторівневого злиття.

сіб EPNet++ досягає стійкішого представлення ознак. RCBEV [?] зосереджується на сприйнятті динамічних об'єктів, перекидаючи місток між відмінностями ознак радара та камери; а DVF [?] збагачує представлення у ділянках низької щільності шляхом генерації багатомасштабних щільних воксельних ознак, уникаючи зашумлених 2D-передбачень з мітками 3D-обмежувальних рамок. LoGoNet [?] поєднує глобальне і локальне злиття з динамічною агрегацією ознак задля підвищення точності детектування у складних середовищах.

Деякі сучасні методи застосовують механізми уваги та контрастне навчання для збагачення мультимодального злиття. Наприклад, CAT-Det [?] кодує глобальну контекстуальну інформацію між модальностями за допомогою контрастного навчання. SeaDATE [?] використовує подвійну увагу та контрастне навчання для витягання глибокої семантичної інформації, тоді як CSSA [?] застосовує легке перемикавання каналів та просторову увагу для ефективного злиття. Fusion-Mamba [?] розв'язує задачу вирівнювання через покращену структуру Mamba та простір прихованих станів.

IV. Методи мультиагентного злиття

У складних відкритих середовищах, особливо за обмеженої видимості або несприятливої погоди, система сприйняття одного втіленого агента стикається з численними викликами. Технологія спільного (колаборативного) сприйняття може інтегрувати дані сприйняття з кількох агентів та інфраструктури, що є вирішальним для подолання проблем перекриттів та відмов сенсорів. У цьому розділі ми зосереджуємось на багатовидовому злитті в режимі агент-до-агента (agent-to-agent, A2A) спільного сприйняття. Рис. ?? показує простий конвеєр злиття A2A.

CoBEVT [?] — перший універсальний мультиагентний мультикамерний каркас сприйняття. Він генерує BEV-передбачення сегментації

через розріджені трансформери для спільного оброблення. CoBEVT містить модуль осьової уваги (axial attention) для ефективного злиття мультиагентних мультивидових ознак камер, охоплюючи як локальні, так і глобальні просторові взаємодії. CoCa3D [?] пропонує інноваційний спільний каркас з виключно камерним сприйняттям. Він розв'язує проблему упередженості передбачення глибини, дозволяючи кільком агентам, обладнаним лише камерами, ділитися візуальною інформацією. Через спільне використання інформації про глибину в однакових точках CoCa3D зменшує помилки, краще опрацьовує невизначеності глибини та розширює можливості детектування на зайняті та далекодійні зони, які зазвичай складні для одноагентних систем. V2VNet [?] вводить каркас на основі графових нейронних мереж для злиття проміжних представлень ознак від кількох транспортних засобів. MACP [?] досліджує ефективну адаптацію моделей з використанням попередньо натренованих одноагентних моделей задля досягнення спільного сприйняття з малою кількістю параметрів та комунікаційних витрат. HM-ViT [?] пропонує уніфікований каркас для проблеми мультимодального A2A-сприйняття, здатний зливати мультимодові ознаки зображень та хмар точок LiDAR з різних типів сенсорів, забезпечуючи ефективне мультимодальне кооперативне сприйняття. MRCNet [?] вирішує проблеми розмиття руху (motion blur), вводячи механізм підсилення руху, який зменшує вплив розмиття руху на детектування об'єктів через захоплення контексту руху, досягаючи кращих результатів у зашумлених сценаріях.

Окрім того, деякі роботи зосереджуються на покращенні комунікаційних аспектів спільного сприйняття задля ефективнішої та стійкішої кооперації. When2Com [?] запропонував каркас навчання тому, як формувати комунікаційні групи і коли комунікувати. Використовуючи механізми "рукостискання" (handshake) та асиметричні розміри повідомлень, він зменшує використання пропускну здатності та досягає хороших результатів у завданнях семантичної сегментації та розпізнавання 3D-форм. Who2Com [?] підвищує точність у завданнях семантичної сегментації через навчання механізмів handshake-комунікації та використовує меншу пропускну здатність порівняно з централізованими методами. How2Com [?] додатково запропонував комунікаційний механізм на основі теорії інформації та просторово-часовий спільний трансформер, що покращує спільне сприйняття через фільтрування ознак, компенсацію затримок та просторово-часове злиття, що дає ефективнішу та стійкішу ко-

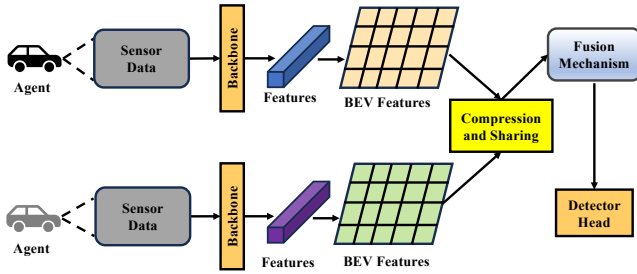


Рис. 6. Простий конвеєр злиття агент-до-агента (A2A).

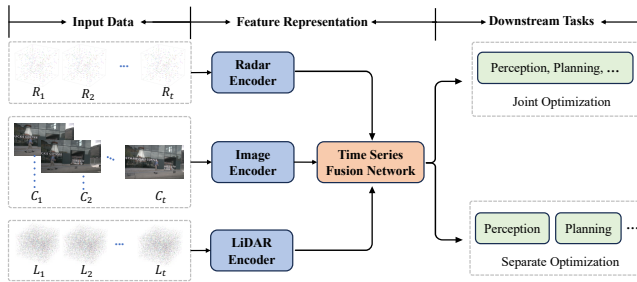


Рис. 7. Загальний огляд каркасу мережі для злиття часових рядів багатосенсорних даних.

операцію у задачах 3D-детектування об'єктів. CodeFilling [?] ефективно оптимізує представлення та вибір спільних повідомлень через стратегію заповнення інформацією та техніки кодокнижкового стиснення, досягаючи ефективного спільного сприйняття з низькими комунікаційними витратами.

V. Злиття часових рядів

Злиття часових рядів є критичним складним систем MSFP: воно долає обмеження окремого кадру та збагачує перцептивну неперервність у просторово-часових доменах. Рис. ?? показує простий конвеєр такого злиття. З появою трансформерних архітектур у комп'ютерному зорі переважними стали методи на основі запитів (query-based), де ознаки сприйняття, закодовані як запити, взаємодіють з просторово-часовими ключами та значеннями для досягнення ефективного вирівнювання ознак. Як показано в табл. ??, ці методи можна класифікувати на три основні категорії: *щільні запити*, *розріджені запити* та *гібридні запити*. Рис. ?? зображує часову лінію query-методів для злиття часових рядів багатосенсорних даних.

V-A. Методи щільних запитів

Методи щільних запитів призначають кожній точці-запиту фіксовану rasterизовану просторову позицію у 3D-просторі високої роздільної здатності або BEV-просторі [?]. Серед них

Табл. IV
Класифікація методів злиття часових рядів на основі запитів.

Категорія	Особливість	Методи
Щільні запити	Зберігають фіксовані просторові позиції у просторах представлення високої роздільної здатності	BEVFormer [?], BEVFormer v2 [?]
Розріджені запити	Ефективно концентрують обчислювальні ресурси на областях інтересу	StreamPETR [?], Sparse4D (v1 [?]/v2 [?]/v3 [?]), SparseFusion3D [?]
Гібридні запити	Поєднують щільну розріджену парадигми	UniAD [?], FusionAD [?], RCBEVDet [?]

особливо репрезентативними є щільні каркаси на основі BEV; BEVFormer [?] став основоположною моделлю BEV-сприйняття. На основі DETR [?] і Deformable DETR [?], BEVFormer досягає адаптивної взаємодії ознак з кількох камерних видів через механізми деформівної уваги. На відміну від декодера у DETR3D [?], який покладається на розріджені об'єктні запити, BEVFormer включає додатковий енкодер на основі щільних BEV-запитів для генерування щільних BEV-ознак, що полегшує задачі семантичної сегментації. BEVFormer зливає часову інформацію між моментами $t - 1$ і t через модуль Temporal Self-Attention у своєму енкодері, що можна інтерпретувати як реалізацію механізму деформівної уваги [?]. Розвиваючи цю основу, BEVFormer v2 [?] приймає двостадійну архітектуру детектування, що інтегрує перспективне детектування з BEV-детектуванням. Це дає BEVFormer v2 змогу навчатися 3D-представленням сцени адаптивно через перспективне нагляд, без покладання на дорогі дані для попереднього навчання глибини.

На основі LSS [?], представника глибинно-орієнтованих “знизу-вгору” підходів, BEVDet4D [?] розширює 3D-детектування на 4D-часовий домен. BEVDet [?] дотримується парадигми LSS і пропонує каркас 3D-детектування мультимодових камер у BEV. Окрім того, BEVDet4D зберігає BEV-ознаки попереднього кадру і зливає їх з ознаками поточного кадру через просторове вирівнювання та конкатенацію. Для подолання впливу власного руху транспортного засобу автори запропонували метод компенсації его-руху на основі згорток і забезпечили точність вирівнювання ознак через допоміжне завдання. Як уніфікований каркас сприйняття і прогнозування, BEVerse [?] генерує 4D-BEV-представлення з мультикамерних відеопослідовностей через спільні модулі екстракції та підняття ознак.

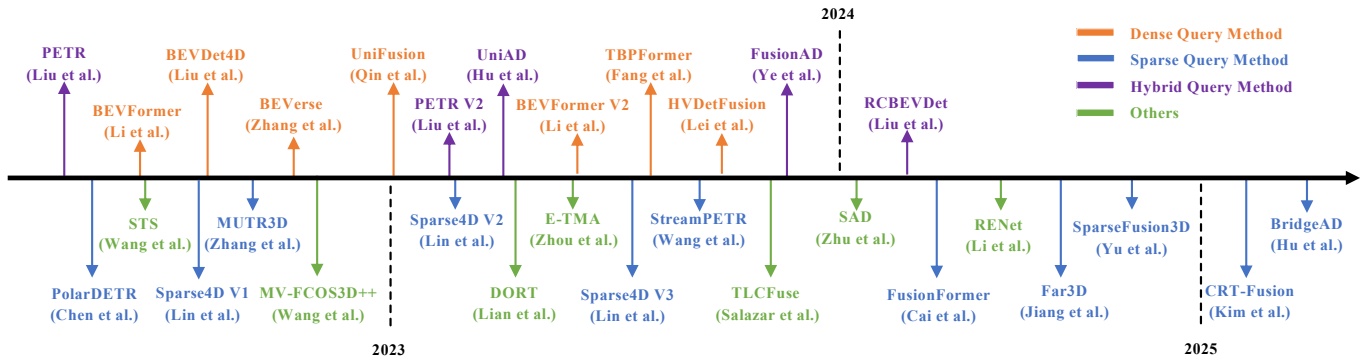


Рис. 8. Часова лінія методів злиття часових рядів.

Існують також методи з унікальними архітектурними рішеннями або оптимізаціями під конкретні завдання. На основі [?], [?], [?], [?], UniFusion [?] пропонує уніфікований каркас часово-просторового злиття, вводячи поняття віртуальних видів. Історичні кадри розглядаються як додаткові камерні види з просторово-перетворювальними відношеннями, що полегшує паралельну обробку часової та просторової інформації. На цій базі TBPFormer [?] зосереджується на загальнішому архітектурному дизайні, пропонуючи енкодер PoseSync BEV для вирівнювання та синхронізації ознак і проектуючи часовий пірамідальний трансформер для багатомасштабної екстракції ознак і прогнозування майбутнього стану. HVDetFusion [?] додатково підтримує радарну модальність і проектує двостадійну розчеплену архітектуру детектування на основі BEVDet4D [?]. Він використовує послідовність із 16 кадрів (*тобто* 8 історичних і 8 майбутніх) для екстракції та злиття ознак, суттєво підвищуючи точність детектування та оцінки швидкості рухомих об'єктів.

V-B. Методи розріджених запитів

У складних відкритих середовищах, коли часова інформація інтегрується у процес злиття, обсяг даних, які мережа обробляє, зростає. Крім того, для задач, що вимагають прийняття рішень у реальному часі, на швидкість виводу моделі мають накладатися суворі вимоги. Тому методи розріджених запитів, відомі своєю ефективністю, точністю та придатністю до задач розрідженого сприйняття, набули поширення в індустрії. У методах щільного BEV-представлення проста деформація багатокандрових BEV-ознак до поточного кадру і їх конкатенація часто дає хороші результати. Однак у представленнях з розрідженими запитами таке явне злиття стає надзвичайно складним. Як наслідок, багато підходів вдаються до взаємодії кожної ознаки-запиту з багатокандровими озна-

ками зображень, що додає значні обчислювальні витрати. StreamPETR [?] розв'язує це шляхом систематичного поширення довгострокової інформації між кадрами через об'єктні запити. Ця об'єктно-центрична парадигма часового моделювання уникає обчислювальних навантажень моделювання часових відношень у щільних BEV-ознаках.

Услід за StreamPETR подальші роботи досягли подальших покращень у представленні ознак та стратегіях вибірки. Однак BEV-методи мають внутрішні компроміси між дальністю сприйняття, точністю та обчислювальною ефективністю, водночас не можучи безпосередньо виконувати 2D-завдання сприйняття у домені зображень. На ці обмеження Sparse4D v1 [?] відповідає, досягаючи ефективної просторово-часової екстракції ознак через 4D-семплінг ключових точок та ієрархічне злиття ознак. На основі Sparse4D v1, Sparse4D v2 [?] застосовує рекурентний підхід, використовуючи розріджені екземпляри для поширення часової інформації, уникаючи багатокандрової вибірки заради ефективнішого злиття ознак. Розвиваючи це, Sparse4D v3 [?] робить наступний крок, пропонуючи часове очищення екземплярів та оцінювання якості, що одночасно прискорює збіжність моделі та підвищує продуктивність.

Багатозадачне навчання відіграє важливу роль у різних задачах сприйвду. Спільне навчання кількох завдань може робити нейромережу громіздкою, але використання розріджених запитів пропонує елегантне рішення. MUTR3D [?] — перший наскрізний 3D-каркас багатокандрового трекінгу, що пов'язує детектування цілей з низхідними завданнями, як-от планування шляху та передбачення траєкторії, через 3D MOT, і пропонує механізм 3D track query, який може моделювати просторово-часову узгодженість цілей між кадрами. На основі MUTR3D, PF-Track [?] приймає каркас "трекінг за допомогою уваги" і узгоджено представляє треківані екземпляри

в часі за допомогою об'єктних запитів. У випадку довготривалих перекриттів PF-Track підтримує позиції об'єктів і дозволяє повторну асоціацію через модуль Future Reasoning, який обробляє історичну інформацію та передбачає стійкі майбутні траєкторії на термін до 4 секунд.

Крім того, нещодавні дослідження демонструють тенденцію до вивчення нових парадигм розрідженого мультимодального часового злиття. Більш ранні методи, як FusionFormer [?], зосереджувались на часовому злитті BEV-ознак для 3D-детектування об'єктів, використовуючи механізми деформівної уваги та залишкові структури для вирівнювання та злиття ознак. Попри інтуїтивну привабливість методів на основі щільних BEV-ознак, більшість з них суттєво втрачає інформацію при обробленні інформації по осі Z. Також на основі DETR [?], QTNNet [?] запропонував нову парадигму часового злиття, що використовує розріджені запити. Модуль моделювання часу, керованого рухом (motion-guided timing modeling, MTM), ефективно опрацьовує крос-модальну кореляцію між хмарами точок та ознаками зображень, досягаючи кращої продуктивності при збереженні легкої архітектури. SparseFusion3D [?] розвиває цей підхід, вводячи модуль MSPCP для передбачення зсуву хмари точок та інтегруючи стратегію ініціалізації запитів за допомогою радару для подолання виклику розрідженості. Еволюція від MTM (QTNNet) до MSPCP (SparseFusion3D) представляє технологічний зсув від простого вирівнювання ознак до явного моделювання на основі руху. Модуль SQS у SparseFusion3D представляє розвиток стратегій мультимодального злиття від простої конкатенації ознак до складніших стратегій, як-от адаптивне зважене злиття. Крім того, CRT-Fusion [?] вирішує задачу врахування руху об'єктів у часовому злитті camera-radar шляхом введення багатокрокових запитів руху, що розрізняють кожен майбутній часовий крок. Метод використовує оцінювач ознак руху (Motion Feature Estimator) для передбачення піксельної швидкості та модуль Motion Guided Temporal Fusion для рекурентного вирівнювання ознак між часовими мітками, досягаючи переваг через явне врахування динаміки об'єктів.

V-C. Методи гібридних запитів

Методи гібридних запитів поєднують парадигми щільних і розріджених запитів, балансуючи обчислювальну ефективність із комплексним розумінням сцени. Ці підходи стратегічно використовують розріджені запити для

Табл. V
Класифікація методів злиття MM-LLM для MSFP.

Категорія	Особливість	Методи
Візуально-мовні	Поєднують візуальні та текстові дані для семантичного вирівнювання.	Scedrivex [?], X-Driver [?], Mpdriive [?], SafeAuto [?]
Візуально-LiDAR-мовні	Інтегрують візуальні, LiDAR-та мовні дані для 3D-просторового розуміння.	DriveMLM [?], MAPLM [?], LiDAR-LLM [?]

об'єктних задач, водночас зберігаючи щільні представлення для просторово-повних задач, досягаючи оптимальної продуктивності у багаточільових задачах сприйняття.

UniAD [?] ілюструє цю гібридну архітектуру, інтегруючи сприйняття, прогнозування та планування в уніфікованому каркасі. Він використовує розріджені об'єктні запити для ефективного детектування і трекінгу, водночас зберігаючи щільні BEV-ознаки для прогнозування траєкторій і завдань планування. Це подвійне представлення забезпечує комплексне розуміння сцени без жертви продуктивності в реальному часі. Розвиваючи успіх UniAD, FusionAD [?] розширює гібридний підхід на мультимодальне часове злиття. Він обробляє дані камер та LiDAR через архітектуру на основі трансформерів, яка адаптивно перемикається між розрідженими та щільними представленнями залежно від вимог завдання, демонструючи гнучкість гібридних методів запитів у роботі з гетерогенними сенсорними даними.

Мультимодальні методи гібридних запитів, що ефективно опрацьовують гетерогенні дані з кількох сенсорів (різні кругові відеокамери або мультимодальні сенсори, *напр.*, 4D-радар міліметрового діапазону, LiDAR, камера) через вишукані архітектури, демонструють чудові здатності у часово-просторовій екстракції та злитті ознак. Спираючись на CRN [?], RCBEVdet [?] вводить двопотокову мережу. Для радарного потоку проектується RadarBEVNet, що генерує щільні BEV-ознаки, для екстракції BEV-ознак з хмар точок. Для камерного потоку використовується image backbone та трансформер виду з LSS [?] для представлення ознак. Далі, через модуль багаторівневого злиття на основі перехресної уваги (на базі deformable DETR), можна краще здійснити ефективне злиття 4D-радар міліметрового діапазону з камерою.

VI. Методи злиття MM-LLM

В останні роки великі мовні моделі (large language model, LLM) досягли вражаючої продуктивності у різноманітних задачах. Зливаючи

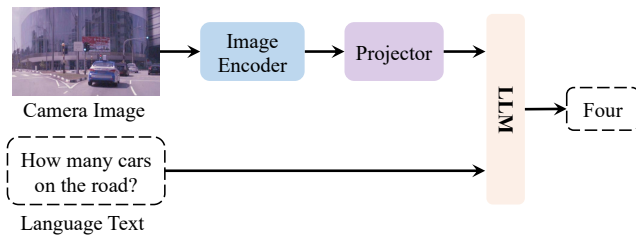


Рис. 9. Парадигма на основі “візуально-мовних” методів.

чи дані з різних модальностей, мультимодальні LLM (MM-LLM) можуть виконувати складніші завдання, як-от описи зображень, розуміння відео та крос-модальний пошук. Останнім часом розроблено різноманітні нові набори даних задля просування MM-LLM для втіленого ШІ. Наприклад, такі проекти, як DriveLM [?], OmniDrive [?] та NuInstruct, збагачують наявні набори даних, включаючи LLM для генерації пар запитання-відповідь, що покривають сприйняття, міркування та планування. Крім того, MAPLM [?] інтегрує мультимодальні зображення з даними LiDAR для аналізу та інтерпретації станів дорожнього покриття. На основі MM-LLM і цих наборів даних проведено багато досліджень щодо інтеграції MM-LLM до MSFP. Як показано в табл. ??, у цьому розділі ми переважно розглядаємо існуючі споріднені роботи з двох категорій: *методи на основі візуально-мовних моделей* та *методи на основі візуально-LiDAR-мовних моделей*. Конвеєри цих двох категорій ілюструють відповідно рис. ?? та рис. ??.

VI-A. Методи на основі візуально-мовних моделей

Мультимодальні великі моделі демонструють значний потенціал у завданнях інтелектуального сприйняття; різні підходи досліджують їхні можливості для розв’язання складнощів реальних середовищ. X-Driver [?] пропонує уніфікований каркас, який використовує мультимодальні великі мовні моделі з ланцюговим міркуванням (Chain-of-Thought) та авторегресійним моделюванням, досягаючи кращої продуктивності для автономного водіння у замкненому контурі та посиленої інтерпретованості. Mprdrive [?] вводить новий каркас підказкового навчання на основі маркерів, що використовує лаконічні візуальні маркери для представлення просторових координат і будує двозерністі візуальні підказки, досягаючи передових результатів зі збагаченням просторовим сприйняттям у задачах, що вимагають розвинутого просторового розуміння. DriveVLM [?] інтегрує традиційні архітектури з MM-LLM через дві окремі

гілки: одна зосереджена на традиційній візуальній обробці, інша використовує потужність мультимодальних трансформерів для розуміння сцени.

Поступ у проектуванні моделей надалі підвищує здатності сприйняття та міркування. Reason2Drive [?] використовує апіорний токенизатор для витягання локальних ознак зображення; BEV-InMLLM [?] вводить BEV-представлення для просторового розуміння; OmniDrive [?] інтегрує 2D попередньо натреновані знання з 3D-просторовими даними через Q-Former3D. Тим часом ELM [?] захоплює часову інформацію через механізм вибору токенів, що враховує час. Крім того, Chen *et al.* [?] пропонують нову архітектуру, що зливає об’єктно-рівневі векторизовані числові модальності в будь-який LLM за допомогою двостадійного методу попереднього тренування та тонкого налаштування.

VI-B. Методи на основі візуально-LiDAR-мовних моделей

Через обмежену доступність даних LiDAR і текстових даних, безпосереднє вирівнювання ознак хмари точок з ознаками тексту є істотним викликом. Ця складність виникає тому, що дані хмари точок, що внутрішньо тривимірні й розріджені, не мають щільної та структурованої природи текстових даних. Щоб подолати ці виклики, ознаки зображення зазвичай використовуються як посередник для ефективного перекидання містка між текстом і LiDAR-даними. Таким чином, багату візуальну інформацію, доступну в зображеннях, можна використати для безшовнішої інтеграції цих різномірних типів даних. У цьому напрямку, DriveMLM [?] застосовує часовий QFormer для обробки мультимодальних зображень, що дозволяє ефективно захоплювати часову динаміку та просторові зв’язки між різними перспективами. Це необхідно для розуміння складних сцен.

Крім того, у мультимодальній обробці деякі методи застосовують непрямий підхід до роботи з хмарами точок: вони перетворюють хмари точок на зображення для полегшення витягання інформації. Таке перетворення дозволяє використати усталені техніки, що відзначаються в обробці зображень, тим самим підвищуючи загальну ефективність MSFP. Наприклад, MAPLM [?] проектує 3D-хмару точок LiDAR на BEV-зображення, а ознаки витягає через візуальний енкодер. Цей підхід перетворює 3D-дані на 2D-представлення, що полегшує обробку традиційними моделями глибокого навчання, призначеними для зображень. Використовуючи BEV-зображення, MAPLM перекидає міст

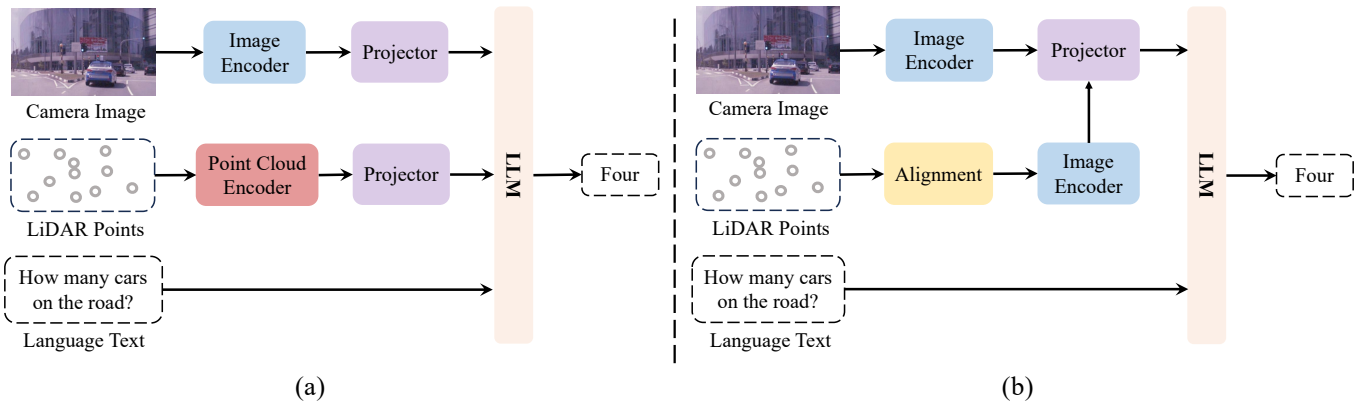


Рис. 10. Парадигми на основі “візуально-LiDAR-мовних” моделей. У парадигмі (a) проєктуються окремі енкодери для радара і зображення, потім відбувається їх злиття. Парадигма (б) зливає радарну хмару точок із зображенням після вирівнювання модальностей зображення.

між даними хмар точок і даними зображень, дозволяючи застосовувати потужні візуальні моделі, як-от CLIP. Крім того, LiDAR-LLM [?] вводить нову схему розуміння 3D-сцен на відкритому повітрі, переформулюючи 3D-когніцію як задачу мовного моделювання; використовується позиційно-усвідомлений трансформер (Position-Aware Transformer, PAT) та триетапна стратегія тренування для подолання модальної прогалини 3D-мова та досягнення передових результатів у задачах 3D-описів, прив'язки та відповідей на запитання.

VII. Відкриті виклики та майбутні можливості

У цьому розділі ми описуємо властиві виклики та можливі майбутні можливості для MSFP. Обговорення нижче ведеться на трьох взаємопов'язаних рівнях: даних, моделі та застосування.

VII-A. Рівень даних

VII-A1. Якість даних: Основний виклик на рівні даних — обмежена якість та репрезентативність наявних наборів даних для багато-сенсорного злиття. Багато наборів даних, як-от KITTI, nuScenes та Waymo Open, страждають на “довгохвостові” розподіли [?], [?], [?]. Така незбалансованість обмежує здатність моделей злиття узагальнюватися до рідкісних, але критичних сценаріїв. Більш того, попри збільшення різноманітності сенсорів, розрідженіші радарні хмари точок також ставлять виклик здатностям існуючих моделей до екстракції ознак. Це поглиблюється такими проблемами, як відсутні дані, викиди, зсуви (bias) і дрейф, а також відсутність стандартизованих методів оцінювання та обмежена доступність публічних наборів даних [?].

Тому розвиток високоякісних наборів даних є істотним для розв'язання цих викликів. Техніки AIGC (Artificial Intelligence Generated Content) потенційно можуть генерувати синтетичні дані для заповнення прогалин у наборах реальних даних, особливо для рідкісних або різноманітних сценаріїв, *напр.*, фотореалістичний рендеринг [?] і дифузійні моделі [?]. Для підвищення надійності згенерованих синтетичних даних майбутні дослідження могли б зосередитися на розробці автоматизованих інструментів виявлення помилок [?]. Крім того, впровадження кількісних метрик якості може допомогти виявляти проблеми, як-от відсутні дані, викиди та дрейф даних.

VII-A2. Аугментація даних: Аугментація даних відіграє життєво важливу роль у підвищенні стійкості та узагальнюваності систем MSFP. Однак мультимодальна аугментація даних також вносить унікальні виклики, зокрема в збереженні синхронізації між різними сенсорними модальностями [?], [?]. Беручи злиття LiDAR-камера як приклад: при застосуванні поворотів або зсувів до хмари точок LiDAR еквівалентні перетворення необхідно застосувати й до відповідних зображень камери для збереження просторової узгодженості [?]. Будь-які невідповідності в перетвореннях можуть зруйнувати просторові відношення між модальностями, що є вирішальним для ефективного злиття сенсорів.

Щоб подолати ці виклики, дослідники могли б зосередитися на розробці передових технік синхронізованої аугментації даних, адаптованих до систем MSFP. Один із потенційних напрямків досліджень — використання крос-модальних геометричних обмежень для забезпечення просторової узгодженості під час аугментації [?]. Наприклад, перетворення хмари точок LiDAR можна поєднати з гомографічними перетвореннями у зображеннях каме-

ри, щоб зберегти точні просторові відношення. Інший перспективний напрямок — AIGC, як-от дифузійні моделі [?], які можуть створювати реалістичні й синхронізовані аугментації для моделювання таких варіацій, як шум сенсорів і зміни середовища, при збереженні крос-модальної узгодженості.

VII-B. Рівень моделі

VII-B1. Ефективні стратегії злиття: На рівні моделі розробка ефективних стратегій злиття передбачає подолання втрати інформації під час вирівнювання та інтеграції мультимодальних сенсорних даних. Втрата інформації часто виникає в процесі вирівнювання через невідповідності між сенсорними модальностями — наприклад, камери та LiDAR — які відрізняються фізичною конфігурацією, роздільною здатністю та перспективою [?]. Фактори, як-от погода і освітлення, посилюють ці розбіжності, ускладнюючи точну синхронізацію [?]. Традиційні методи вирівнювання, як-от проектування хмар точок у систему координат камери [?], часто вносять додаткові помилки, що ведуть до субоптимальної інтеграції. Процеси інтеграції, як-от агрегація ознак, також додають до втрати інформації, стискаючи сенсорні дані та упускаючи критичні деталі. Наприклад, перетворення хмар точок у 2D-проекції (як-от BEV або range view) зменшує тривимірну просторову інформацію, як-от висоту, що життєво важлива для захоплення геометрії сцени [?], [?]. Ця сукупна втрата зменшує здатність моделі повноцінно використати взаємодоповнювальні переваги різних сенсорів, обмежуючи загальну продуктивність.

Щоб пом'якшити ці виклики, майбутні дослідження могли б зосередитися на стратегіях злиття, що спільно оптимізують точність вирівнювання та злиття даних. Техніки злиття з кількома представленнями, як-от поєднання воксельних ґраток, хмар точок і 2D-проекцій, відкривають шлях до збереження просторового та семантичного багатства [?]. Контекстно-усвідомлені підходи, що використовують часову узгодженість [?], та методи адаптивного навчання [?] можуть покращити вирівнювання через динамічну реакцію на зміни середовища. Крім того, механізми уваги [?] можуть вибірково підкреслювати критичні ознаки кожної модальності під час інтеграції. Більше того, такі техніки, як самоконтрольоване представницьке навчання [?] і контрастне навчання [?], мають потенціал захоплювати й використовувати крос-модальні зв'язки, надаючи багатшого й детальнішого нагляду для уточнення точності вирівнювання. Ці рішення є основою для

зменшення втрати інформації та підвищення стійкості мультимодальних систем злиття.

VII-B2. Підхід на основі MM-LLM: MM-LLM можна використати для обробки та злиття даних з різноманітних джерел — тексту, зображень і сенсорних виходів — що може значно збагатити розуміння складних середовищ [?], [?]. Однак інтеграція цих моделей у реальні застосування втіленого ШІ все ще ставить критичні виклики. Істотний виклик — робота з розрідженими та нерегулярними сенсорними даними, як-от хмари точок LiDAR і радара. Висока вимірність даних радарних хмар точок вимагає вишуканих технік попередньої обробки та екстракції ознак для перетворення їх на формат, придатний для входу моделі. Крім того, радарні дані за своєю природою розріджені й неструктуровані, що ускладнює їх обробку порівняно зі структурованішими типами даних, як-от зображення чи текст [?], [?]. Щоб подолати цю прогалину, майбутні дослідження могли б досліджувати гібридні архітектури, які поєднують техніки геометричного навчання — як-от графові нейромережі або моделі точкового навчання [?], [?] — з мультимодальними здатностями MM-LLM.

Крім того, зовнішні знання MM-LLM, навчених на різноманітних наборах даних, можуть конфліктувати зі специфічними вимогами втіленого ШІ. Наприклад, MM-LLM може зробити висновок, що сценарій із завантаженим пішохідним переходом застосовний універсально, пропонуючи зайві зупинки в середовищах, як-от шосе, де пішохідні переходи нерелевантні. Такі конфлікти між загальними знаннями та специфічними контекстами можуть скомпрометувати прийняття рішень, якщо ними обережно не керувати або не адаптувати їх. Щоб подолати ці виклики, майбутні дослідження могли б досліджувати такі механізми, як Retrieval-Augmented Generation (RAG), для динамічної адаптації зовнішніх знань до контексту, наданого багатосенсорними даними [?]. Механізми уваги можуть додатково уточнювати цей процес, акцентуючи релевантну інформацію та відфільтровуючи нерелевантний або вводячий в оману контент. Ці підходи надають потенційний шлях до забезпечення того, щоб зовнішні знання узгоджувалися зі специфічними реальними вимогами систем втілених агентів, підвищуючи їхню стійкість і надійність.

VII-C. Рівень застосування

VII-C1. Адаптивність до реального світу: У реальних відкритих середовищах умови суттєво змінюються: змінюються освітлення, погода та схеми руху. Раптові зміни — дощ, туман, сніг

або перехід від дня до ночі — становлять істотні виклики для мультимодальних систем злиття, які повинні постійно підтримувати надійну продуктивність попри ці динамічні зміни. Тому ефективна адаптація до різноманітних сценаріїв є істотною для забезпечення стійкості систем MSFP і запобігання відмовам у складних умовах.

Щоб покращити адаптивність до реального світу, майбутні дослідження могли б зосередитися на розробці самоадаптивних алгоритмів, що можуть налаштовувати параметри моделі у відповідь на зміни середовища в реальному часі [?]. Такі техніки, як адаптація доменів та онлайн-навчання [?], [?], дозволяють моделям підтримувати продуктивність, постійно адаптуючись до нових розподілів даних без перетренування з нуля. Крім того, можна досліджувати методи zero-shot learning [?], [?], що дозволяють моделям узагальнюватися до невидимих сценаріїв і опрацьовувати нові умови середовища без попереднього специфічного тренування, поліпшуючи їхню здатність впоратися з непередбачуваними реальними ситуаціями.

VII-C2. Інтерпретованість (Explainability): Інтерпретованість суттєво важлива для моделей MSFP у втіленому III, оскільки вона будує довіру, забезпечує прозорість і допомагає в налагодженні, особливо в безпеково-критичних застосуваннях [?]. Один з основних викликів полягає у виявленні внеску кожної сенсорної модальності у різних умовах. Наприклад, LiDAR може грати ключову роль у поганому освітленні, тоді як камери можуть бути ефективнішими за ясної погоди. Розуміння цих внесків — а також того, як різні модальності взаємодіють — складне, оскільки ці відношення часто не прямі, особливо у складних реальних сценаріях.

Майбутні дослідження могли б досліджувати методи контекстно-усвідомленої інтерпретованості [?] для прояснення ролі кожної модальності залежно від умов середовища та стадій злиття. Наприклад, інструменти візуалізації на основі уваги [?], [?] можуть підкреслити, які сенсори зробили найбільший внесок у конкретних сценаріях, підвищуючи прозорість прийняття рішень. Крім того, можна проектувати інтерпретовані мережі злиття, що видають специфічні для модальності оцінки впевненості, надаючи чітке розуміння того, як кожне джерело даних вплинуло на вихід, особливо в критичних або неоднозначних ситуаціях. Адаптація пояснень до модальностей та сценаріїв застосування може посилити довіру користувача та підтримати безпечніше й ефективніше розгортання у реальних середовищах.

VIII. Висновок

У цьому огляді ми всебічно розглянули методи досліджень багатосенсорного сприйняття на основі злиття даних (MSFP) для втіленого III. Зокрема, ми спочатку ввели передумови MSFP. Далі впорядкували й розглянули конкретні методи за чотирма категоріями: методи мультимодального злиття, мультиагентного злиття, злиття часових рядів та злиття MM-LLM. Нарешті, обговорили поточні виклики та майбутні можливості. Озираючись назад: на відміну від існуючих оглядів, що зосереджуються на конкретних областях (напр., автономне керування) або задачах (напр., 3D-детектування об'єктів), ми впорядкували дослідження MSFP з незалежної від завдання перспективи, де методи представлено суто з різних технічних точок зору. Тому ця стаття придатна для прочитання дуже широкого кола дослідників, у різних областях і з різних задач. У майбутньому ми регулярно оновлюватимемо цей огляд онлайн, щоб подавати найсвіжіший передовий поступ у галузі MSFP.

З огляду на обмежену експертизу авторів, якщо у статті є якісь недоліки, просимо повідомляти. Ми серйозно покращимо огляд у наступній версії.

Література

- [1] S. Ruan, Y. Zhang, K. Zhang, Y. Fan, F. Tang, Q. Liu, and E. Chen, "Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 960–13 969.
- [2] S. Ruan, H. Liu, Z. Chen, B. Feng, K. Zhang, C. C. Cao, E. Chen, and L. Chen, "Cpws: Confident programmatic weak supervision for high-quality data labeling," *ACM Transactions on Information Systems*, vol. 43, no. 4, pp. 1–26, 2025.
- [3] S. Ruan, K. Zhang, L. Wu, T. Xu, Q. Liu, and E. Chen, "Color enhanced cross correlation net for image sentiment analysis," *IEEE Transactions on Multimedia*, vol. 26, pp. 4097–4109, 2024.
- [4] L. Ren, J. Dong, S. Liu, L. Zhang, and L. Wang, "Embodied intelligence toward future smart manufacturing in the era of ai foundation model," *IEEE/ASME Transactions on Mechatronics*, 2024.
- [5] A. Gupta, S. Savarese, S. Ganguli, and L. Fei-Fei, "Embodied intelligence via learning and evolution," *Nature communications*, vol. 12, no. 1, p. 5721, 2021.
- [6] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, "Multi-modal 3d object detection in autonomous driving: a survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 2122–2152, 2023.
- [7] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Tang, L. Yang, J. Li, C. Jia et al., "Multi-modal 3d object detection in autonomous driving: A survey and taxonomy," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 7, pp. 3781–3798, 2023.
- [8] C. Xiang, C. Feng, X. Xie, B. Shi, H. Lu, Y. Lv, M. Yang, and Z. Niu, "Multi-sensor fusion and cooperative perception for autonomous driving: A review," *IEEE Intelligent Transportation Systems Magazine*, 2023.
- [9] J. Zhu, H. Li, and T. Zhang, "Camera, lidar, and imu based multi-sensor fusion slam: A survey," *Tsinghua Science and Technology*, vol. 29, no. 2, pp. 415–429, 2023.

- [10] Y. Tang, H. He, Y. Wang, Z. Mao, and H. Wang, "Multi-modality 3d object detection in autonomous driving: A review," *Neurocomputing*, p. 126587, 2023.
- [11] H. Du, L. Ren, Y. Wang, X. Cao, and C. Sun, "Advancements in perception system with multi-sensor fusion for embodied agents," *Information Fusion*, p. 102859, 2024.
- [12] T. Bin, H. Yan, N. Wang, M. N. Nikolić, J. Yao, and T. Zhang, "A survey on the visual perception of humanoid robot," *Biomimetic Intelligence and Robotics*, p. 100197, 2024.
- [13] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, and L. Wang, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 11, pp. 15407–15436, 2024.
- [14] X. Han, S. Chen, Z. Fu, Z. Feng, L. Fan, D. An, C. Wang, L. Guo, W. Meng, X. Zhang *et al.*, "Multimodal fusion and vision-language models: A survey for robot vision," *arXiv preprint arXiv:2504.02477*, 2025.
- [15] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.
- [16] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11621–11631.
- [17] P. Sun, H. Kretschmar, A. Dotiwalla, C. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.
- [18] N. Gährlert, N. Jourdan, M. Cordts, U. Franke, and J. Denzler, "Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.
- [19] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, and D. Ramanan, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.
- [20] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A*3d: An autonomous driving dataset in challenging environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 994–995.
- [21] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, R. Yang, Z. Carmichael, C. Lang *et al.*, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.
- [22] X. Weng, Y. Man, D. Cheng, J. Park, M. O'Toole, and K. Kitani, "All-in-one drive: A large-scale comprehensive perception dataset with high-density long-range point clouds," *arXiv preprint arXiv:2010.03180*, 2020.
- [23] S. Patil, F. P'atzold, C. H'ane, M. Tschentscher, and A. Knoll, "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9552–9558.
- [24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, H. Dahlkamp, A. Schuster, U. Franke, and S. Roth, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223. [Online]. Available: <https://arxiv.org/abs/1604.01685>
- [25] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 244–253.
- [26] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.
- [27] T. Yin, X. Zhou, and P. Krähnenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16494–16507, 2021.
- [28] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17182–17191.
- [29] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.
- [30] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11794–11803.
- [31] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18442–18455, 2022.
- [32] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3d detection with depth completion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5418–5427.
- [33] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.
- [34] K. Shin, Y. P. Kwon, and M. Tomizuka, "Roarnet: A robust 3d object detection based on region approximation refinement," in *2019 IEEE intelligent vehicles symposium (IV)*. IEEE, 2019, pp. 2510–2515.
- [35] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5127–5137.
- [36] J. Zhang, M. Cao, X. Yang, W. Xie, J. Lei, D. Li, W. Huang, and Y. Li, "E2e-mfd: Towards end-to-end synchronous multimodal fusion detection," *arXiv preprint arXiv:2403.09323*, 2024.
- [37] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.
- [38] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, "Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1523–1535, 2023.
- [39] K. Zhou, L. Chen, and X. Cao, "Improving multispectral pedestrian detection by addressing modality imbalance problems," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 787–803.
- [40] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 403–411.
- [41] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

- [42] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [43] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.
- [44] L. Xie, C. Xiang, Z. Yu, G. Xu, Z. Yang, D. Cai, and X. He, "Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12460–12467.
- [45] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, "Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3047–3054.
- [46] Z. Song, H. Wei, L. Bai, L. Yang, and C. Jia, "Graphalign: Enhancing accurate feature alignment by graph matching for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3358–3369.
- [47] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 5291–5304, 2022.
- [48] Y. Li, X. Qi, Y. Chen, L. Wang, Z. Li, J. Sun, and J. Jia, "Voxel field fusion for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1120–1129.
- [49] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao, "Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection," *arXiv preprint arXiv:2201.06493*, 2022.
- [50] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Deformable feature aggregation for dynamic multi-modal 3d object detection," in *European conference on computer vision*. Springer, 2022, pp. 628–644.
- [51] Z. Song, G. Zhang, J. Xie, L. Liu, C. Jia, S. Xu, and Z. Wang, "Voxelnextfusion: A simple, unified and effective voxel fusion framework for multi-modal 3d object detection," *arXiv preprint arXiv:2401.02702*, 2024.
- [52] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.
- [53] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5363–5371.
- [54] H. Zhang, E. Fromont, S. Lefèvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 72–80.
- [55] T. Zhao, M. Yuan, F. Jiang, N. Wang, and X. Wei, "Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion," *arXiv preprint arXiv:2401.10731*, 2024.
- [56] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.
- [57] M. Zhu, C. Ma, P. Ji, and X. Yang, "Cross-modality 3d object detection," in *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, 2021, pp. 3772–3781.
- [58] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7345–7353.
- [59] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV* 16. Springer, 2020, pp. 35–52.
- [60] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, and X. Bai, "Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 7, pp. 8324–8341, 2022.
- [61] A. Mahmoud, J. S. Hu, and S. L. Waslander, "Dense voxel fusion for 3d object detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 663–672.
- [62] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao et al., "Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17524–17534.
- [63] Y. Zhang, J. Chen, and D. Huang, "Cat-det: Contrastively augmented transformer for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 908–917.
- [64] S. Dong, W. Xie, D. Yang, J. Tian, Y. Li, J. Zhang, and J. Lei, "Seadate: Remedy dual-attention transformer with semantic alignment via contrast learning for multimodal object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [65] W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, X. Liu, J. Zhang, G. Guo, and B. Zhang, "Fusion-mamba for cross-modality object detection," *arXiv preprint arXiv:2404.09146*, 2024.
- [66] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.
- [67] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake lidar in 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9243–9252.
- [68] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2vnet: Vehicle-to-vehicle communication for joint perception and prediction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II* 16. Springer, 2020, pp. 605–621.
- [69] Y. Ma, J. Lu, C. Cui, S. Zhao, X. Cao, W. Ye, and Z. Wang, "Macp: Efficient model adaptation for cooperative perception," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3373–3382.
- [70] H. Xiang, R. Xu, and J. Ma, "Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 284–295.
- [71] S. Hong, Y. Liu, Z. Li, S. Li, and Y. He, "Multi-agent collaborative perception via motion-aware robust communication network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15301–15310.
- [72] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.
- [73] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.

- [74] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [75] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 481–15 490.
- [76] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [77] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 830–17 839.
- [78] S. Wang, Y. Liu, T. Wang, Y. Li, and X. Zhang, "Exploring object-centric temporal modeling for efficient multi-view 3d object detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3598–3608.
- [79] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion," 2022.
- [80] X. Lin, T. Lin, Z. Pei, L. Huang, and Zhizhong Su, "Sparse4d v2: Recurrent temporal fusion with sparse model," *arXiv preprint arXiv:2305.14018*, 2023.
- [81] X. Lin, Z. Pei, T. Lin, L. Huang, and Z. Su, "Sparse4d v3: Advancing end-to-end 3d detection and tracking," 2023.
- [82] Z. Yu, W. Wan, M. Ren, X. Zheng, and Z. Fang, "Sparsefusion3d: Sparse sensor fusion for 3d object detection by radar and camera in environmental perception," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1524–1536, 2024.
- [83] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [84] T. Ye, W. Jing, C. Hu, S. Huang, L. Gao, F. Li, J. Wang, K. Guo, W. Xiao, W. Mao, H. Zheng, K. Li, J. Chen, and K. Yu, "Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving," *arXiv preprint arXiv:2308.01006*, 2023.
- [85] Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, and C. Zhu, "Rcbvnet: Radar-camera fusion in bird's eye view for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 928–14 937.
- [86] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, and X. Zhu, "Vision-centric bev perception: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 978–10 997, 2024.
- [87] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [88] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=gZ9hCDWe6ke>
- [89] Y. Wang, V. Guizilini, T. Zhang, Y. Wang, H. Zhao, , and J. M. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *The Conference on Robot Learning (CoRL)*, 2021.
- [90] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly estimating depth," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [91] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.
- [92] J. Huang, G. Huang, Z. Zhu, Y. Yun, D. Du, and X. Bai, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [93] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.
- [94] Z. Qin, J. Chen, C. Chen, X. Chen, and X. Li, "Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8656–8665.
- [95] S. Fang, Z. Wang, Y. Zhong, J. Ge, and S. Chen, "Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1368–1378.
- [96] K. C. Lei, Z. Chen, S. Jia, and X. Zhang, "Hvdtetfusion: A simple and robust camera-radar fusion framework," *arXiv preprint arXiv:2307.11323*, 2023.
- [97] T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, "Mutr3d: A multi-camera tracking framework via 3d-to-2d queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4537–4546.
- [98] Z. Pang, J. Li, P. Tokmakov, D. Chen, S. Zagoruyko, and Y.-X. Wang, "Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [99] Y. Cai, W. Zhang, Y. Wu, and C. Jin, "Fusionformer: A concise unified feature fusion transformer for 3D pose estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [100] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.
- [101] J. Hou, Z. Liu, Z. Zou, X. Ye, X. Bai *et al.*, "Query-based temporal fusion with explicit motion for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 75 782–75 797, 2023.
- [102] J. Kim, M. Seong, and J. W. Choi, "Crt-fusion: Camera, radar, temporal fusion using motion information for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 108 625–108 648, 2024.
- [103] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862.
- [104] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "Crn: Camera radar net for accurate, robust, efficient 3d perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [105] R. Zhao, Q. Yuan, J. Li, H. Hu, Y. Li, C. Zheng, and F. Gao, "Sce2drivex: A generalized mllm framework for scene-to-drive learning," *arXiv preprint arXiv:2502.14917*, 2025.
- [106] W. Liu, J. Zhang, B. Zheng, Y. Hu, Y. Lin, and Z. Zeng, "X-driver: Explainable autonomous driving with vision-language models," *arXiv preprint arXiv:2505.05098*, 2025.
- [107] Z. Zhang, X. Li, Z. Xu, W. Peng, Z. Zhou, M. Shi, and S. Huang, "Mpdrive: Improving spatial understanding with marker-based prompt learning for autonomous driving," *arXiv preprint arXiv:2504.00379*, 2025.

- [108] J. Zhang, X. Yang, T. Wang, Y. Yao, A. Petiushko, and B. Li, "Safeauto: Knowledge-enhanced safe autonomous driving with multimodal foundation models," *arXiv preprint arXiv:2503.00211*, 2025.
- [109] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.
- [110] X. Cao, T. Zhou, Y. Ma, W. Ye, C. Cui, K. Tang, Z. Cao, K. Liang, Z. Wang, J. M. Rehg *et al.*, "Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 819–21 830.
- [111] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, H. Li, Y. Guo *et al.*, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9247–9255.
- [112] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," *arXiv preprint arXiv:2312.14150*, 2023.
- [113] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.
- [114] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.
- [115] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," in *European Conference on Computer Vision*. Springer, 2025, pp. 292–308.
- [116] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 668–13 677.
- [117] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied understanding of driving scenarios," in *European Conference on Computer Vision*. Springer, 2025, pp. 129–148.
- [118] L. Chen, O. Sinavski, J. Hünemann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 093–14 100.
- [119] H. Y. Teh, A. W. Kempa-Liehr, and K. I.-K. Wang, "Sensor data quality: A systematic review," *Journal of Big Data*, vol. 7, no. 1, p. 11, 2020.
- [120] Z. Yang, H. Yang, Z. Pan, and L. Zhang, "Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting," in *The Twelfth International Conference on Learning Representations*.
- [121] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [122] L. Ehrlinger and W. Wöß, "A survey of data quality measurement and monitoring tools," *Frontiers in big data*, vol. 5, p. 850611, 2022.
- [123] A. Rangesh, N. Deo, R. Greer, P. Gunaratne, and M. M. Trivedi, "Predicting take-over time for autonomous driving with real-world data: Robust data augmentation, models, and evaluation," *arXiv preprint arXiv:2107.12932*, 2021.
- [124] A. Xiao, J. Huang, D. Guan, K. Cui, S. Lu, and L. Shao, "Polarmix: A general data augmentation technique for lidar point clouds," *Advances in Neural Information Processing Systems*, vol. 35, pp. 11 035–11 048, 2022.
- [125] X. Yuan, Y. Xie, S. Wang, and T. Xiong, "Camera-lidar extrinsic calibration via traffic signs," in *2023 42nd Chinese Control Conference (CCC)*. IEEE, 2023, pp. 4679–4684.
- [126] G. P. Meyer, J. Charland, D. Hegde, A. Laddha, and C. Vallespi-Gonzalez, "Sensor fusion for joint 3d object detection and semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.
- [127] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023.
- [128] H. Zhong, H. Wang, Z. Wu, C. Zhang, Y. Zheng, and T. Tang, "A survey of lidar and camera fusion enhancement," *Procedia Computer Science*, vol. 183, pp. 579–588, 2021.
- [129] J. Zhao, J. Shi, and L. Zhuo, "Bev perception for autonomous driving: State of the art and future perspectives," *Expert Systems with Applications*, vol. 258, p. 125103, 2024.
- [130] C. Chang, J. Zhang, K. Zhang, W. Zhong, X. Peng, S. Li, and L. Li, "Bev-v2x: Cooperative birds-eye-view fusion and grid occupancy prediction via v2x-based data sharing," *IEEE Transactions on Intelligent Vehicles*, 2023.
- [131] Y. Xu, X. Tong, and U. Stilla, "Voxel-based representation of 3d point clouds: Methods, applications, and its potential use in the construction industry," *Automation in Construction*, vol. 126, p. 103675, 2021.
- [132] L. Zhu, W. Yang, B. Chen, H. Zhu, X. Meng, and S. Wang, "Temporally consistent enhancement of low-light videos via spatial-temporal compatible learning," *International Journal of Computer Vision*, pp. 1–21, 2024.
- [133] A. Lübken and A. Förster, "An adaptive approach to time synchronization for wireless sensors under extreme conditions," *IEEE Sensors Journal*, 2024.
- [134] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.
- [135] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 2022.
- [136] K. Wickstrøm, M. Kampffmeyer, K. Ø. Mikalsen, and R. Jenssen, "Mixing up contrastive learning: Self-supervised representation learning for time series," *Pattern Recognition Letters*, vol. 155, pp. 54–61, 2022.
- [137] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 040–13 051.
- [138] T. Tian, B. Li, X. Weng, Y. Chen, E. Schmerling, Y. Wang, B. Ivanovic, and M. Pavone, "Tokenize the world into object-level knowledge to address long-tail events in autonomous driving," in *8th Annual Conference on Robot Learning*.
- [139] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wöhler, "Radarscenes: A real-world radar point cloud data set for automotive applications," in *2021 IEEE 24th International Conference on Information Fusion (FUSION)*. IEEE, 2021, pp. 1–8.
- [140] Y. Li, L. Ma, Z. Zhong, F. Liu, M. A. Chapman, D. Cao, and J. Li, "Deep learning for lidar point clouds in autonomous driving: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3412–3432, 2020.
- [141] P. Svenningsson, F. Fioranelli, and A. Yarovsky, "Radar-pointgcn: Graph based object recognition for unstructured radar point-cloud data," in *2021 IEEE*

- Radar Conference (RadarConf21)*. IEEE, 2021, pp. 1–6.
- [142] B. A. Murray, N. C. Coops, L. Winiwarter, J. C. White, A. Dick, I. Barbeito, and A. Ragab, “Estimating tree species composition from airborne laser scanning data using point-based deep learning models,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 207, pp. 282–297, 2024.
 - [143] S. Sharifmoghaddam, S. Upadhyay, W. Chen, and J. Lin, “Unirag: Universal retrieval augmentation for multi-modal large language models,” *arXiv preprint arXiv:2405.10311*, 2024.
 - [144] J. Xing, D. Wei, S. Zhou, T. Wang, Y. Huang, and H. Chen, “A comprehensive study on self-learning methods and implications to autonomous driving,” *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
 - [145] M. Silva, T. Medeiros, M. Azevedo, M. Medeiros, M. Themoteo, T. Gois, I. Silva, and D. G. Costa, “An adaptive tinyml unsupervised online learning algorithm for driver behavior analysis,” in *2023 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*. IEEE, 2023, pp. 199–204.
 - [146] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, “Online learning: A comprehensive survey,” *Neurocomputing*, vol. 459, pp. 249–289, 2021.
 - [147] W. Wang, V. W. Zheng, H. Yu, and C. Miao, “A survey of zero-shot learning: Settings, methods, and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.
 - [148] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, “Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2251–2265, 2018.
 - [149] K. P. De Jong Yeong and J. Walsh, “Building trustworthy autonomous vehicles: The role of multi-sensor fusion and explainable ai (xai) in on-road and off-road scenarios,” *Sensors and Electronic Instrumentation Advances*, p. 145, 2024.
 - [150] J. Wang and M. Bilgic, “Context-aware feature selection and classification,” in *IJCAI*, 2023, pp. 4317–4325.
 - [151] S. Jin, H. Lee, C. Park, H. Chu, Y. Tae, J. Choo, and S. Ko, “A visual analytics system for improving attention-based traffic forecasting models,” *IEEE transactions on visualization and computer graphics*, vol. 29, no. 1, pp. 1102–1112, 2022.
 - [152] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viégas, and M. Wattenberg, “Attentionviz: A global view of transformer attention,” *IEEE Transactions on Visualization and Computer Graphics*, 2023.