دانشگاه خواجه نصیرالدین طوسی
دانشکده مهندسی کامپیوتر- گروه هوش مصنوعی

پایان‌نامه دوره کارشناسی ارشد

مهندسی کامپیوتر

تخمین عمق بدون نظارت با استفاده از پارامترهای کالیبراسیون داخلی به عنوان ورودی

محمدامین قاسم‌زاده

استاد راهنما

دکتر بهروز نصیحت کن

تابستان ۱۴۰۵

دانشگاه خواجه نصیرالدین طوسی
دانشکده مهندسی کامپیوتر- گروه هوش مصنوعی

پایان‌نامه دوره کارشناسی ارشد

مهندسی کامپیوتر

عنوان

تخمین عمق بدون نظارت با استفاده از پارامترهای کالیبراسیون داخلی به عنوان ورودی

نگارش

محمدامین قاسم‌زاده

استاد راهنما

دکتر بهروز نصیحت کن

تابستان ۱۴۰۵

بسم الله الرحمن الرحیم

تقدیم به: مادر و پدر عزیزم

به آنان که با علم خود زندگی آزاد می‌سازند

دانشگاه صنعتی خواجه نصیرالدین طوسی

تأییدیهٔ هیئت داوران جلسه‌ی دفاع از پایان‌نامهٔ کارشناسی ارشد

هیأت داوران پس از مطالعه‌ی پایان‌نامهٔ و شرکت در جلسه‌ی دفاع از پایان‌نامهٔ تهیه‌شده با عنوان «تخمین عمق بدون نظارت با استفاده از پارامترهای کالیبراسیون داخلی به عنوان ورودی» توسط آقای / خانم محمدامین قاسم‌زاده صحت و کفایت تحقیق انجام شده را برای اخذ درجه‌ی کارشناسی ارشد در رشته‌ی مهندسی کامپیوتر در تاریخ تابستان ۱۴۰۵مورد تأیید قرار دادند.

۱. استاد راهنما: دکتر بهروز نصیحت کن . . . . . امضا

۲. استاد داور داخلی: دکتر حمید ابریشمی مقدم. . . . . امضا

۳. استاد مدعو: دکتر محمدرضا محمدی . . . . . امضا

دانشگاه صنعتی خواجه نصیرالدین طوسی

اظهارنامه دانشجو

اینجانب محمدامین قاسم‌زاده به شماره دانشجویی ۴۰۲۰۹۶۴۴دانشجوی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر دانشکده دانشکدهٔ مهندسی کامپیوتر دانشگاه‌خواجه نصیرالدین طوسی گواهی می‌نمایم که تحقیقات ارائه شده در این پایان‌نامه با عنوان:

تخمین عمق بدون نظارت با استفاده از پارامترهای کالیبراسیون داخلی به عنوان ورودی

توسط اینجانب انجام و بدون هرگونه دخل و تصرف است و موارد نسخه برداری شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده ام. در صورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم (قانون حمایت از حقوق مؤلفان و مصنفان و قانون ترجمه و تکثیر کتب و نشریات و آثار صوتی، ضوابط و مقررات آموزشی، پژوهشی و انضباطی و غیره) با اینجانب رفتار خواهد شد. در ضمن، مسئولیت هرگونه پاسخگویی به اشخاص اعم از حقیقی و حقوقی و مراجع ذی صلاح (اعم از اداری و قضایی) به عهده ی اینجانب خواهد بود و دانشگاه هیچ گونه مسئولیتی در این خصوص نخواهد داشت.

نام و نام خانوادگی دانشجو: محمدامین قاسم‌زاده

تاریخ و امضای دانشجو:

دانشگاه صنعتی خواجه نصیرالدین طوسی

حق طبع، نشر و مالکیت نتایج

حق چاپ و تکثیر این پایان نامه متعلق به نویسندگان آن می باشد. بهره برداری از این پایان نامه در چهارچوب مقررات کتابخانه و با توجه به محدودیتی که توسط استاد راهنما به شرح زیر تعیین می گردد، بلامانع است:

- ☐ بهره‌برداری از این پایان‌نامه برای همگان و با ذکر منبع، بلامانع است.
- ☐ بهره‌برداری از این پایان‌نامه با اخذ مجوز از استاد راهنما و با ذکر منبع، بلامانع است.
- ☐ بهره‌برداری از این پایان‌نامه تا تاریخ _____ ممنوع است.

استاد راهنما: دکتر بهروز نصیحت کن . . . . . امضا

قدردانی

اکنون که به یاری پروردگار و یاری و راهنمایی اساتید بزرگ موفق به پایان این رساله شده‌ام وظیفه خود دانشته که نهایت سپاسگزاری را از تمامی عزیزانی که در این راه به من کمک کرده‌اند را به عمل آورم: در آغاز از استاد بزرگ و دانشمند جناب آقادکتر نصیحت‌کن که راهنمایی این پایان‌نامه را به عهده داشته‌اند کمال تشکر را دارم. از داوران گرامی دکتر حمید ابریشمی مقدم و دکتر محمدرضا محمدی که زحمت داوری و تصحیح این پایان‌نامه را به عهده داشتند کمال سپاس را دارم. خالصانه از تمامی اساتید و معلمان و مدرسانی که در مقاطع مختلف تحصیلی به من علم آموخته و مرا از سرچشمه دانایی سیراب کرده‌اند متشکرم. از کلیه هم دانشگاهیان و همراهان عزیز، دوستان خوبم نهایت سپاس را دارم.

و در پایان این پایان‌نامه را تقدیم می‌کنم به مادر و پدر مهربانم که با حضور و همراهی شان در دوران تحصیل، همیشه راه را به من نشان داده و مرا در این راه استوار و ثابت قدم نموده است.

محمدامین قاسم‌زاده

تابستان ۱۴۰۵

چکیده

تخمین عمق تک‌دوربینه ذاتاً مسئله‌ای مبهم است، زیرا یک تصویر منفرد مقیاس متریک صحنه را به‌صورت یکتا تعیین نمی‌کند. در این پایان‌نامه بررسی می‌شود که آیا ورود صریح پارامترهای درونی دوربین می‌تواند این ابهام را در یک مدل بنیادین تخمین عمق کاهش دهد. برای این منظور، نسخه‌ای آگاه از دوربین از مدل Depth Anything V2 توسعه داده شده است که در آن یک توصیف‌گر نه‌بعدی نرمال‌شده از ماتریس مشخصهٔ دوربین، به توکنی دروازه‌بانی‌شده تبدیل و به انکودر DINOv2 تزریق می‌شود. روش تجربی طی مجموعه‌ای از آزمایش‌های کنترل‌شده تکامل یافت و در این مسیر، نشت داده در تفکیک NYUv2، ناهم‌خوانی خاموش میان فضای خروجی دیسپاریتی و هدف عمق مستقیم، به‌روزرسانی‌نشدن K هنگام برش تصویر و ناهم‌خوانی خاموش سر خروجی مدل متریک در ارزیابی شناسایی و اصلاح شد.

تمام مقایسه‌های نهایی بر مجموعهٔ آزمون بدون نشت ۶۵۴ تصویری Eigen و میان دو بازوی آموزشی همسان، مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K)، انجام شده‌اند. در پروتکل نسبی، مدل پیشنهادی (دارای K)، نسخهٔ ViT-S، از مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده و مدل کنترل (فاقد K) بهتر است؛ با این حال، در نسخهٔ ViT-L سهم خالص K معنادار نیست، زیرا تراز مقیاس و جابجایی به‌ازای هر تصویر، مهم‌ترین اطلاعات حاصل از فاصلهٔ کانونی را حذف می‌کند. در ارزیابی متریک بدون تراز، مدل پیشنهادی (دارای K) به‌طور پایدار از مدل کنترل (فاقد K) بهتر عمل می‌کند و با افزایش تنوع میدان دید، اندازهٔ اثر نیز افزایش می‌یابد. سپس افزونه‌سازی هندسی «لرزش کانونی» پیشنهاد می‌شود که بازیابی مقیاس را تنها از روی پیکسل‌ها ناممکن می‌سازد. در این تنظیم، مدل پیشنهادی (دارای K) به AbsRel برابر ۰٫۰۸۴۹ در برابر ۰٫۱۰۳۰ برای مدل کنترل (فاقد K) می‌رسد و در همهٔ معیارهای گزارش‌شده با $p \leq ۱۰^{-۲۴}$ برتری دارد.

آزمون‌های سازوکاری و علّی نشان می‌دهند که این بهبود واقعاً ناشی از استفاده از ماتریس مشخصه است. شیب لگاریتمی پاسخ عمق پیش‌بینی‌شده به مقیاس فاصلهٔ کانونی، پس از لرزش کانونی از ۰٫۰۶ به ۰٫۹۳ تا ۰٫۹۷ می‌رسد که به مقدار ایده‌آل یک در مدل روزنه‌ای نزدیک است. حذف K خطای AbsRel را ۵۹ درصد افزایش می‌دهد و تحریف عمدی فاصلهٔ کانونی، خطایی هم‌اندازه با پیش‌بینی هندسهٔ تصویربرداری ایجاد می‌کند. بنابراین، پارامترهای درونی دوربین می‌توانند سیگنالی معنادار و از نظر علّی ضروری برای بازیابی مقیاس فراهم کنند؛ هرچند این اتکا، حساسیت به دقت کالیبراسیون را نیز افزایش می‌دهد و تعمیم آن به داده‌های واقعی چنددوربینه و محیط‌های بیرونی نیازمند بررسی‌های آینده است.

واژگان کلیدی تخمین عمق تک‌دوربینه، یادگیری عمیق بدون نظارت، عمق متریک، پارامترهای درونی دوربین، هندسهٔ تصویربرداری، ۲V Anything Depth

فهرست مطالب

فهرست تصاویر چ

فهرست جداول خ

فهرست الگوریتم‌ها ر

فهرست برنامه‌ها ژ

فهرست اختصارات س

فصل ۱: مقدمه ۱

۱.۱ بیان مسئله و اهمیت آن ۱

۲.۱ چالش‌های عمق متریک در مدل‌های یادگیری عمیق ۳

۳.۱ خلأ پژوهشی ۳

۴.۱ اهداف و سوالات تحقیق ۴

۵.۱ ساختار پایان‌نامه ۵

فصل ۲: مروری بر مطالعات انجام شده ۹

۱.۲ مقدمه ۹

۲.۲ روش‌های کلاسیک تخمین عمق تک‌دوربینه ۱۰

۳.۲ روش‌های مبتنی بر یادگیری عمیق ۱۱

۴.۲ تخمین عمق متریک و غیرمتریک ۱۲

۵.۲ نقش هندسهٔ دوربین و اطلاعات درونی ۱۴
۶.۲ جمع‌بندی فصل و جایگاه پژوهش حاضر ۱۴

فصل ۳: روش تحقیق ۱۹
۱.۳ مقدمه ۱۹
۲.۳ صورت‌بندی مسئلهٔ تخمین عمق تک‌دوربینه ۲۰
۳.۳ هندسهٔ دوربین و نقش پارامترهای درونی ۲۲
۱.۳.۳ مدل دوربین سوراخ‌سوزنی ۲۲
۲.۳.۳ ماتریس مشخصهٔ دوربین ۲۳
۳.۳.۳ نقش پارامترهای درونی در تخمین عمق ۲۴
۴.۳ معماری پایه: مدل Depth Anything V2 . . . . . ۲۵
۱.۴.۳ ساختار کلی شبکهٔ encoder-decoder ۲۵
۲.۴.۳ بخش encoder مبتنی بر Transformer Vision ۲۶
۳.۴.۳ بخش decoder و راهبرد بازسازی نگاشت عمق ۲۶
۵.۳ انگیزه‌ی صریح برای لحاظ پارامترهای درونی دوربین ۲۷
۱.۵.۳ شواهد تجربی از تاثیر پارامترهای دوربین در معیارهای متریک ۲۷
۲.۵.۳ عدم کفایت رویکرد ضمنی نسبت به مشخصه‌های دوربین ۲۸
۳.۵.۳ کاربردهای عملی و انتقال‌پذیری به داده‌های جهان واقعی ۳۰
۴.۵.۳ لزوم طراحی مدل با ورود صریح پارامترهای درونی ۳۰
۶.۳ طراحی روش پیشنهادی: ادغام صریح ماتریس مشخصهٔ دوربین ۳۰
۱.۶.۳ اصل طراحی: تفکیک اطلاعات تصویری و هندسی ۳۱
۲.۶.۳ نمایش و پیش‌پردازش ماتریس مشخصهٔ دوربین ۳۱
۳.۶.۳ راهبرد ادغام اطلاعات دوربین با معماری Depth Anything V2 . . . . . ۳۲
۴.۶.۳ مزایای مورد انتظار از ادغام صریح پارامترهای درونی ۳۴
۷.۳ راهبرد آموزش مدل پیشنهادی (دارای K) ۳۵
۱.۷.۳ مجموعه‌داده‌های مورد استفاده ۳۶

۲.۷.۳ تفکیک صحیح مجموعه‌های آموزش و آزمون در ۲NYUv ۳۷
۳.۷.۳ مدل پایهٔ Depth Anything V2 و راهبرد فاین‌تیون ۳۷
۴.۷.۳ فعال‌سازی پشتیبانی از پارامترهای دوربین ۳۸
۵.۷.۳ تنظیمات آموزش و بهینه‌سازی ۳۸
۶.۷.۳ فضای هدف آموزش برای مدل‌های نسبی: عمق معکوس ۳۹
۷.۷.۳ افزونه‌سازی هندسی آگاه از پارامترهای دوربین ۴۰
۸.۷.۳ افزونه‌سازی لرزش کانونی: غیرقابل‌حل‌کردن مقیاس بدون K ۴۲
۹.۷.۳ استفاده از تقطیر دانش ۴۴
۱۰.۷.۳ جمع‌بندی فرآیند آموزش ۴۴
۸.۳ توابع هزینه و معیارهای بهینه‌سازی ۴۴
۱.۸.۳ هزینهٔ SiLog برای مدل‌های متریک ۴۵
۲.۸.۳ هزینهٔ تغییرناپذیر نسبت به مقیاس و جابجایی برای مدل‌های نسبی ۴۶
۳.۸.۳ تمایز میان توابع هزینهٔ آموزش و معیارهای ارزیابی ۴۷
۱.۳.۸.۳ توابع هزینه در آموزش ۴۷
۲.۳.۸.۳ معیارهای ارزیابی ۴۷
۴.۸.۳ نقش پارامترهای درونی دوربین در آموزش ۴۸
۵.۸.۳ حساسیت RMSE به پارامترهای دوربین و کاربرد آن در ارزیابی ۴۹
۶.۸.۳ جمع‌بندی ۴۹
۹.۳ جزئیات پیاده‌سازی ۵۰
۱.۹.۳ چارچوب نرم‌افزاری و زیرساخت محاسباتی ۵۰
۲.۹.۳ پیش‌پردازش داده‌ها ۵۰
۳.۹.۳ مدیریت و نرمال‌سازی پارامترهای درونی دوربین ۵۱
۴.۹.۳ تنظیمات آموزش و راهبرد بهینه‌سازی ۵۲
۵.۹.۳ دیتاست‌های مورد استفاده در پیاده‌سازی ۵۲
۶.۹.۳ ساختار کد و اسکریپت‌های اصلی ۵۴
۷.۹.۳ ملاحظات عملی و محدودیت‌ها ۵۶

۸.۹.۳ اشکال‌زدایی و درس‌های پیاده‌سازی ۵۷
۹.۹.۳ جمع‌بندی ۵۸
۱۰.۳ جمع‌بندی فصل ۵۸

فصل ۴: نتایج ۶۱
۱.۴ مقدمه ۶۱
۲.۴ سوالات تحقیق و فرضیات تجربی ۶۲
۱.۲.۴ بازگویی سوالات تحقیق ۶۲
۲.۲.۴ فرضیات تجربی ۶۳
۳.۲.۴ ارتباط فرضیات با بخش‌های فصل ۶۳
۳.۴ تنظیمات آزمایش و پروتکل ارزیابی ۶۴
۱.۳.۴ دیتاست‌های مورد استفاده ۶۴
۲.۳.۴ نوع محیط‌ها و سناریوهای ارزیابی ۶۵
۳.۳.۴ معیارهای ارزیابی ۶۶
۱.۳.۳.۴ معیارهای غیرمتریک ۶۶
۲.۳.۳.۴ معیارهای متریک ۶۶
۴.۳.۴ آزمون‌های آماری ۶۶
۴.۴ نتایج کمی: مقایسهٔ عددی مدل‌ها ۶۷
۱.۴.۴ عملکرد مدل پایهٔ Depth Anything V2 در دیتاست‌های مختلف ۶۷
۲.۴.۴ مقایسهٔ مدل پایهٔ Depth Anything V2 و مدل پیشنهادی (دارای K) ۶۸
۳.۴.۴ تحلیل پایداری بین‌دیتاستی مدل پیشنهادی (دارای K) ۶۹
۵.۴ مطالعهٔ کنترل‌شده بر روی پروتکل استاندارد NYUv2 . . . . . ۷۰
۱.۵.۴ ممیزی نقاط وارسی و مسیر رسیدن به آزمایش نهایی ۷۰
۲.۵.۴ راستی‌آزمایی زنجیرهٔ ارزیابی: بازتولید نتایج مدل پایهٔ Depth Anything V2 . . ۷۵
۳.۵.۴ ناکارآمدی فاین‌تیون برای ارزیابی برون‌دامنه‌ای ۷۵
۴.۵.۴ نتایج پروتکل نسبی: مقایسهٔ چهارگانه ۷۶

۵.۵.۴ اثر افزونه‌سازی هندسی و یک نتیجهٔ منفی آموزنده ۷۹
۶.۵.۴ ارزیابی متریک: آشکارشدن سهم پارامترهای درونی دوربین ۸۰
۷.۵.۴ دور سوم — افزونه‌سازی لرزش کانونی: جهش سهم پارامترهای دوربین ۸۲
۸.۵.۴ مطالعهٔ فروکاست: تفکیک اثر لرزش کانونی از ضریب λ ۸۳
۹.۵.۴ نمای یکپارچهٔ چهار زوج آموزشی و الگوی پاسخ به دوز ۸۴
۱۰.۵.۴ جمع‌بندی مطالعهٔ کنترل‌شده و گزارش کوشش‌های تجربی ۸۴
۱۱.۵.۴ نمای اندازهٔ اثر: مقایسهٔ همهٔ معیارها در یک نگاه ۸۶
۶.۴ تحلیل حساسیت نسبت به پارامترهای دوربین ۸۶
۱.۶.۴ آزمون جاروی K: آیا مدل واقعاً فاصلهٔ کانونی را می‌خواند؟ ۸۷
۲.۶.۴ آزمون حذف و تحریف K: اثبات اتکای علّی ۸۸
۳.۶.۴ رفتار متفاوت معیارهای AbsRel و RMSE ۹۰
۴.۶.۴ ارتباط با تحلیل هندسی فصل سوم ۹۰
۷.۴ ملاحظات کیفی و حدود شواهد بصری ۹۱
۸.۴ تحلیل موردی دیتاست CityScapes ۹۱
۱.۸.۴ ماهیت و ویژگی‌های خاص دیتاست CityScapes ۹۳
۲.۸.۴ تحلیل کیفیت عمق مرجع ۹۴
۳.۸.۴ دلیل رفتار متفاوت خطاها ۹۴
۴.۸.۴ محدودیت همسان‌سازی نتایج با سایر دیتاست‌ها ۹۴
۹.۴ بحث و تفسیر نتایج ۹۵
۱.۹.۴ جمع‌بندی تحلیلی نتایج کمی و کیفی ۹۵
۲.۹.۴ پاسخ به سوالات تحقیق ۹۵
۳.۹.۴ داوری صریح فرضیه‌های تجربی ۹۶
۴.۹.۴ وابستگی سهم پارامترهای دوربین به پروتکل ارزیابی و ظرفیت مدل ۹۷
۵.۹.۴ مقایسهٔ غیرمستقیم با ادبیات موضوع ۹۷
۶.۹.۴ تفسیر نتایج در چارچوب هندسهٔ تصویربرداری ۹۸
۱۰.۴ جمع‌بندی فصل نتایج ۹۸

فصل ۵: بحث و نتیجه‌گیری ۱۰۱

۱.۵ مقدمه ۱۰۱

۲.۵ جمع‌بندی یافته‌های اصلی پژوهش ۱۰۲

۳.۵ تفسیر نهایی نتایج در چارچوب هندسهٔ تصویربرداری ۱۰۴

۴.۵ نوآوری‌های تحقیق ۱۰۵

۱.۴.۵ نوآوری تئوریک ۱۰۵

۲.۴.۵ نوآوری عملی ۱۰۶

۳.۴.۵ دستاوردهای روش‌شناختی و تشخیصی ۱۰۶

۵.۵ محدودیت‌ها و چالش‌های پژوهش ۱۰۷

۶.۵ پیشنهادهایی برای تحقیقات آینده ۱۱۰

۷.۵ جمع‌بندی نهایی فصل ۱۱۲

کتاب‌نامه ۱۱۳

واژه‌نامهٔ فارسی به انگلیسی **اول**

واژه‌نامهٔ انگلیسی به فارسی **سوم**

فهرست تصاویر

۱.۱ ماهیت هندسی ابهام مقیاس در تخمین عمق تک‌دوربینه ۲

۲.۱ ساختار پایان‌نامه ۷

۱.۲ دسته‌بندی روش‌های تخمین عمق تک‌دوربینه ۱۰

۲.۲ سیر تحول تخمین عمق تک‌دوربینه ۱۲

۳.۲ سه پارادایم نظارتی رایج و جایگاه ماتریس مشخصهٔ دوربین در هر یک ۱۳

۴.۲ چهار راهبرد ممکن برای برخورد با ماتریس مشخصهٔ دوربین ۱۵

۱.۳ مدل هندسی دوربین سوراخ‌سوزنی ۲۳

۲.۳ مقایسهٔ زوجی دو سمت دوربین در Middlebury پس از میانگین‌گیری دو نسخهٔ تصحیح هر یک از ۲۳ محتوای صحنه. ردیف بالا مقادیر هر صحنه و میانگین کل (لوزی سیاه) و ردیف پایین اختلاف مرتب‌شدهٔ دوربین ۰ منهای دوربین ۱ را نشان می‌دهد. خط‌چین قرمز میانگین اختلاف است. ۲۹

۳.۳ معماری مدل پیشنهادی و مسیر تزریق ماتریس مشخصهٔ دوربین ۳۴

۴.۳ مقایسهٔ سه سازوکار ممکن برای شرطی‌سازی مدل بر پارامترهای دوربین ۳۵

۵.۳ افزونه‌سازی هندسی آگاه از پارامترهای دوربین ۴۱

۶.۳ منطق افزونه‌سازی لرزش کانونی ۴۳

۱.۴ نمونه‌ای از تنوع شرایط تصویربرداری مجموعه‌داده‌ها ۶۵

۲.۴ مقایسهٔ بین‌دیتاستی معیار SiLog ۶۹

۳.۴ سه فاز مسیر تجربی پژوهش ۷۵

۴.۴ امضای فروریزش خاموش در منحنی‌های آموزش ۷۷
۵.۴ دو پروتکل ارزیابی و آنچه تراز مقیاس از ارزیابی حذف می‌کند ۸۰
۶.۴ تعمیم به میدان دیدهای متفاوت ۸۳
۷.۴ الگوی پاسخ به دوز در چهار زوج آموزشی مستقل ۸۵
۸.۴ اندازهٔ اثر زوجی همهٔ معیارها با بازهٔ اطمینان ۹۵ درصد ۸۶
۹.۴ آزمون جاروی K: پاسخ مقیاس خروجی به فاصلهٔ کانونی ورودی ۸۸
۱۰.۴ آزمون حذف و تحریف K در برابر پیش‌بینی هندسی ۸۹
۱۱.۴ مقایسهٔ کیفی روی نمونهٔ پیش‌ثبت‌شده ۹۲
۱۲.۴ توزیع نسبت مقیاس هر تصویر ۹۳

۱.۵ نمای کلی از فرآیند و دستاوردهای پژوهش ۱۰۴

فهرست جداول

۱.۱ نگاشت سوالات تحقیق به فرضیه‌ها و شواهد ۶

۱.۲ مقایسهٔ ساختاریافتهٔ رویکردهای شاخص تخمین عمق تک‌دوربینه ۱۶

۲.۲ صورت‌بندی خلأ پژوهشی و پاسخ پژوهش حاضر ۱۷

۱.۳ ممیزی بازتولیدپذیر تفاوت دو سمت دوربین در Middlebury. اختلاف به‌صورت دوربین ۰ منهای دوربین ۱ و بازه‌ها به‌صورت بوت‌استرپ زوجی ۹۵ درصد گزارش شده‌اند ($n = ۲۳$). ۲۸

۲.۳ مؤلفه‌های بردار توصیف‌گر هندسی دوربین ۳۳

۳.۳ مجموعه‌داده‌های به‌کاررفته و نقش آن‌ها در پژوهش ۳۶

۴.۳ جمع‌بندی توابع هزینه، فضای هدف و پروتکل ارزیابی متناظر ۴۸

۵.۳ ابرپارامترهای آموزش مطالعهٔ کنترل‌شده ۵۳

۶.۳ سه اشکال خاموش شناسایی‌شده در زنجیرهٔ آموزش و ارزیابی ۵۸

۱.۴ مقایسهٔ معیار SiLog مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده در دیتاست‌های مختلف ۶۸

۲.۴ مقایسهٔ عملکرد مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده و مدل پیشنهادی (دارای K) بر اساس معیار SiLog ۶۸

۳.۴ مسیر تصمیم‌گیری از نقاط وارسی اولیه تا آزمایش نهایی آگاه از دوربین. این جدول رتبه‌بندی مدل‌ها نیست؛ هر سطر نشان می‌دهد یک مشاهده چگونه به یک برداشت و سپس به تصمیم آزمایشی بعدی منجر شد. ۷۱

۴.۴ راستی‌آزمایی زنجیرهٔ ارزیابی: بازتولید نتایج مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ نسبی، بر روی پروتکل استاندارد. هر معیار در ستونی مستقل گزارش شده است. ۷۵

۵.۴ ارزیابی بدون نشت نسخه‌های فاین‌تیون‌شدهٔ چندمجموعه‌ای در حالت صفرنمونه، در مقایسه با مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ نسبی ViT-S. هر معیار در ستونی مستقل گزارش شده است. ۷۶

۶.۴ مقایسهٔ چهارگانه بر روی مجموعهٔ آزمون استاندارد NYUv2 (۶۵۴ تصویر، مرجع عمق خام)، انکودر ViT-S، پروتکل نسبی با تراز مقیاس و جابجایی. ۷۶

۷.۴ مقایسهٔ چهارگانه بر روی مجموعهٔ آزمون استاندارد NYUv2، انکودر ViT-L، پروتکل نسبی. ۷۸

۸.۴ اندازهٔ اثر زوجی مدل پیشنهادی (دارای K) ViT-S بر ۶۵۴ تصویر NYUv2. . . . . ۷۸

۹.۴ آزمون‌های زوجی معناداری بر روی ۶۵۴ تصویر آزمون NYUv2 (پروتکل نسبی)؛ مقایسهٔ مدل پیشنهادی (دارای K) با مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده و مدل کنترل (فاقد K). ۷۸

۱۰.۴ اثر افزونه‌سازی هندسی آگاه از دوربین بر زوج مدل ViT-L در پروتکل نسبی NYUv2. . ۷۹

۱۱.۴ ارزیابی متریک (عمق مطلق، بدون تراز) بر روی مجموعهٔ آزمون استاندارد NYUv2، انکودر ViT-L. دور ۱: افزونه‌سازی با میدان دید ۴۱-۶۳ درجه؛ دور ۲: ۳۳-۶۳ درجه. . . ۸۱

۱۲.۴ آزمون‌های زوجی معناداری در ارزیابی متریک ($n = ۶۵۴$)؛ مقایسهٔ مدل پیشنهادی (دارای K) با مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ متریک، و مدل کنترل (فاقد K). ۸۱

۱۳.۴ دور ۳ (لرزش کانونی): ارزیابی متریک (عمق مطلق، بدون تراز) بر مجموعهٔ آزمون استاندارد NYUv2، انکودر ViT-L، به‌همراه آزمون زوجی ویلکاکسون ($n = ۶۵۴$) و بازهٔ اطمینان ۹۵٪ بوت‌استرپ تفاضل زوجی. ۸۲

۱۴.۴ مطالعهٔ فروکاست (لرزش کانونی با $\lambda = ۰/۵$ بدون تغییر): ارزیابی متریک بر مجموعهٔ آزمون استاندارد NYUv2 ($n = ۶۵۴$). ۸۳

۱۵.۴ نمای یکپارچهٔ چهار زوج آموزشی مستقل ۸۴

۱۶.۴ شیب لگاریتمی جاروی K (میانگین $\pm$ انحراف معیار روی ۲۰ تصویر آزمون): پاسخ عمق پیش‌بینی‌شده به مقیاس‌دهی مصنوعی فاصلهٔ کانونی. شیب ایده‌آل برای مدل آگاه از هندسه ۱ است. ۸۷

۱۷.۴ آزمون حذف و تحریف K بر مدل پیشنهادی (دارای K) دور ۳ (ViT-L متریک، $n =$ ۶۵۴، بدون تراز). پیش‌بینی نظری: با شیب جاروی $s =$ ۰٫۹۳، تحریف f در ضریب X باید AbsRel را به حدود $|X^s -$ ۱$|$ برساند. ۸۹

۱۸.۴ داوری فرضیه‌های تجربی بر پایهٔ شواهد فصل چهارم ۹۶

۱.۵ جمع‌بندی نوآوری‌ها و دستاوردهای پژوهش ۱۰۸

۲.۵ نگاشت محدودیت‌های پژوهش به مسیرهای پژوهشی آینده ۱۱۰

فهرست الگوریتم‌ها

۱.۳ روند آموزش مدل با پارامترهای هندسی ۴۵

فهرست برنامه‌ها

فهرست اختصارات

فصل ۱

مقدمه

۱.۱ بیان مسئله و اهمیت آن

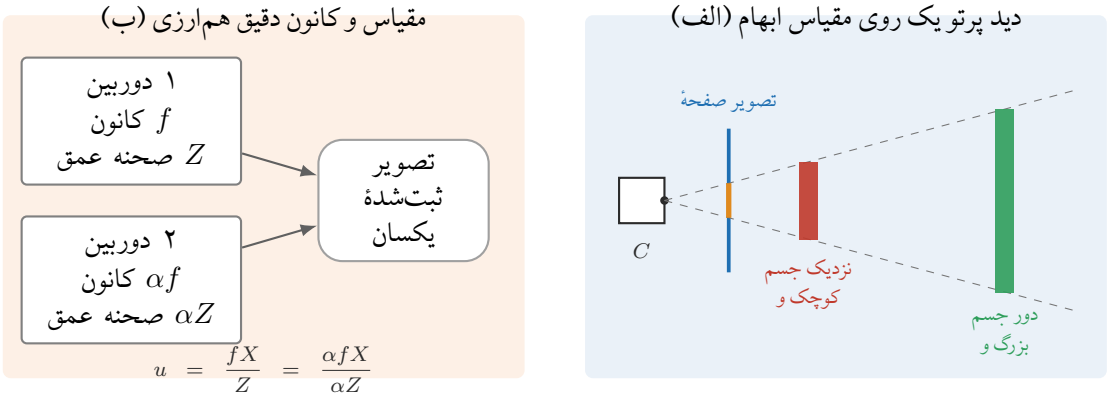
تخمین عمق صحنه از روی یک تصویر تک‌دوربینه یکی از مسائل بنیادی و در عین حال چالش‌برانگیز در بینایی ماشین[1] به شمار می‌رود [۱، ۲]. این مسئله در کاربردهای متنوعی از جمله رباتیک متحرک[2]، ناوبری خودکار[3]، بازسازی سه‌بعدی[4]، واقعیت افزوده[5] و سیستم‌های کمک‌راننده[6] نقش کلیدی ایفا می‌کند [۳]. برخلاف روش‌های مبتنی بر چند نما یا حسگرهای فعال، تخمین عمق تک‌دوربینه[7] تنها بر اطلاعات یک تصویر دوبعدی متکی است و از این‌رو ذاتاً با ابهام‌های هندسی متعددی مواجه می‌شود [۴].

با پیشرفت روش‌های یادگیری عمیق، به‌ویژه شبکه‌های عصبی کانولوشنی[8] و مدل‌های مبتنی بر ترنسفورمر، کیفیت تخمین عمق تک‌دوربینه به‌طور قابل توجهی بهبود یافته است [۵، ۶]. بسیاری از مدل‌های مدرن قادرند ساختار نسبی صحنه را با دقت بالا بازسازی کنند و روابط عمقی میان اجسام را به‌خوبی بیاموزند. با این حال، خروجی این مدل‌ها در اغلب موارد فاقد مقیاس فیزیکی واقعی بوده و به‌صورت عمق نسبی یا *non-metric depth* بیان می‌شود. این محدودیت باعث می‌شود که استفادهٔ مستقیم از چنین مدل‌هایی در کاربردهایی که نیازمند اندازه‌گیری‌های متریک دقیق هستند، با چالش مواجه شود.

---

[1] Computer Vision
[2] Mobile Robotics
[3] Autonomous Navigation
[4] 3D Reconstruction
[5] Augmented Reality
[6] Advanced Driver-Assistance Systems (ADAS)
[7] Monocular Depth Estimation
[8] Convolutional Neural Networks

مسئلهٔ اصلی در این میان، ابهام مقیاس ذاتی در تخمین عمق تک‌دوربینه است. در غیاب اطلاعات هندسی صریح دربارهٔ دوربین تصویربرداری، بازیابی عمق متریک تنها بر اساس شدت پیکسل‌ها ذاتاً نامعین است [۳]. اگرچه برخی روش‌ها با اعمال نرمال‌سازی یا استفاده از معیارهای نسبی، عملکرد مناسبی در سنجش‌های غیرمتریک نشان می‌دهند، اما این رویکردها تضمینی برای پایداری مقیاس فیزیکی عمق فراهم نمی‌کنند. در نتیجه، تغییر شرایط تصویربرداری، از جمله تفاوت در نوع دوربین یا پارامترهای آن، می‌تواند منجر به تغییرات معنادار در کیفیت تخمین عمق شود.



شکل ۱.۱: ماهیت هندسی ابهام مقیاس در تخمین عمق تک‌دوربینه. (الف) دو جسم با اندازه و فاصلهٔ متفاوت که روی یک پرتو دید قرار دارند، تصویر کاملاً یکسانی تولید می‌کنند؛ بنابراین شدت پیکسل‌ها به‌تنهایی مقیاس مطلق صحنه را تعیین نمی‌کند. (ب) صورت کمی همین ابهام: اگر فاصلهٔ کانونی و تمام عمق‌های صحنه هم‌زمان در ضریب α ضرب شوند، تصویر حاصل بدون هیچ تغییری باقی می‌ماند. رفع این ابهام تنها با دانستن فاصلهٔ کانونی — یعنی بخشی از ماتریس مشخصهٔ دوربین K — ممکن است. همین هم‌ارزی، در فصل سوم مبنای افزونه‌سازی «لرزش کانونی» قرار می‌گیرد.

از این منظر، دستیابی به تخمین عمق متریک[1] پایدار و قابل تعمیم، همچنان یکی از چالش‌های باز در حوزهٔ بینایی ماشین محسوب می‌شود. پرداختن به این مسئله نه‌تنها از دیدگاه نظری اهمیت دارد، بلکه از منظر عملی نیز برای توسعهٔ سیستم‌های قابل اعتماد در محیط‌های واقعی ضروری است. این پژوهش با تمرکز بر این چالش اساسی، به بررسی راهکارهایی برای ارتقای مدل‌های تخمین عمق تک‌دوربینه از خروجی‌های غیرمتریک به عمق متریک پایدار می‌پردازد.

[1] Metric Depth

۲.۱ چالش‌های عمق متریک در مدل‌های یادگیری عمیق

اگرچه مدل‌های یادگیری عمیق در سال‌های اخیر پیشرفت چشمگیری در تخمین عمق تک‌دوربینه داشته‌اند، اما اغلب این مدل‌ها به‌طور ضمنی برای یادگیری ساختار نسبی صحنه طراحی شده‌اند [۲]. بسیاری از معیارهای ارزیابی رایج نیز بر سنجش خطاهای نرمال‌شده یا نسبی تمرکز دارند و به‌جای مقیاس فیزیکی مطلق، همبستگی ساختاری عمق را مورد توجه قرار می‌دهند [۶]. این موضوع سبب شده است که بهبود عملکرد در این معیارها لزوماً به معنای دستیابی به عمق متریک دقیق نباشد.

یکی از چالش‌های مهم در این زمینه، وابستگی مدل‌ها به توزیع داده‌های آموزشی است. مدل‌هایی که بر روی یک یا چند دیتاست خاص آموزش داده می‌شوند، معمولاً فرض ضمنی ثبات شرایط تصویربرداری را در خود دارند. در عمل، تغییر نوع دوربین، فاصلهٔ کانونی یا میدان دید[1] می‌تواند توزیع نگاشت تصویر به عمق را تغییر دهد و باعث افت عملکرد مدل شود [۷]. این مسئله به‌ویژه در معیارهای متریک که نسبت به مقیاس فیزیکی حساس هستند، نمود بیشتری پیدا می‌کند.

چالش دیگر، فقدان استفادهٔ صریح از اطلاعات هندسی دوربین در بسیاری از مدل‌های موجود است [۸]. در حالی که فرآیند تصویربرداری به‌طور مستقیم تحت تأثیر پارامترهای درونی دوربین[2] قرار دارد، اغلب روش‌های یادگیری عمیق این اطلاعات را نادیده می‌گیرند یا به‌صورت ضمنی در داده‌ها مستتر فرض می‌کنند. این رویکرد می‌تواند منجر به یادگیری روابطی شود که تنها در شرایط خاص معتبر بوده و قابلیت تعمیم به سناریوهای متفاوت را ندارند.

در مجموع، این چالش‌ها نشان می‌دهند که برای دستیابی به تخمین عمق متریک پایدار، صرف بهبود معماری شبکه یا افزایش حجم داده‌های آموزشی کافی نیست. بلکه لازم است به‌صورت آگاهانه به نقش هندسهٔ تصویربرداری و اطلاعات دوربین در فرآیند تخمین عمق توجه شود. این ضرورت، انگیزهٔ اصلی پژوهش حاضر را شکل می‌دهد و زمینه را برای ارائهٔ رویکردی فراهم می‌کند که بتواند پایداری عمق متریک را در مواجهه با تغییرات شرایط تصویربرداری بهبود بخشد.

۳.۱ خلأ پژوهشی

بخش عمده‌ای از پژوهش‌های پیشین بر بهبود دقت تخمین عمق نسبی تمرکز داشته‌اند [۲، ۹] و موفق شده‌اند ساختار کلی صحنه و روابط عمقی میان اجسام را با کیفیت مناسبی بازسازی کنند. با این حال، در بسیاری از این

[1] Field of View (FoV)
[2] Camera Intrinsics

مطالعات، مسئلهٔ مقیاس فیزیکی عمق یا به‌صورت ضمنی نادیده گرفته شده و یا به‌عنوان یک چالش ثانویه تلقی شده است.

از سوی دیگر، روش‌هایی که به تخمین عمق متریک پرداخته‌اند، غالباً متکی بر فرض‌های محدودکننده‌ای نظیر ثبات نوع دوربین یا شرایط تصویربرداری کنترل‌شده [۶] یا دسترسی به داده‌های کالیبره‌شدهٔ خاص بوده‌اند. این فرض‌ها موجب می‌شود که قابلیت تعمیم چنین مدل‌هایی به سناریوهای واقعی و متنوع کاهش یابد. به‌ویژه، تأثیر تغییر پارامترهای درونی دوربین بر عملکرد مدل‌ها در بسیاری از این کارها به‌صورت نظام‌مند مورد مطالعه قرار نگرفته است.

نکتهٔ قابل توجه دیگر آن است که در اغلب پژوهش‌های موجود، ارزیابی مدل‌ها عمدتاً بر اساس معیارهای غیرمتریک[1] یا نرمال‌شده انجام شده است [۵]. این معیارها اگرچه برای سنجش کیفیت نسبی عمق مناسب هستند، اما قادر به آشکارسازی ناپایداری‌های مرتبط با مقیاس فیزیکی و تغییر شرایط تصویربرداری نیستند. در نتیجه، ممکن است مدلی از نظر این معیارها عملکرد مطلوبی داشته باشد، در حالی که در کاربردهای متریک و واقعی دچار خطای قابل توجه شود.

علاوه بر این، اگرچه نقش هندسهٔ دوربین در فرآیند تصویربرداری به‌خوبی شناخته شده است، استفادهٔ صریح از اطلاعات درونی دوربین در مدل‌های یادگیری عمیق تخمین عمق، کمتر مورد توجه قرار گرفته است [۳، ۸]. در بسیاری از رویکردها، این اطلاعات یا کاملاً نادیده گرفته می‌شوند یا فرض می‌شود که شبکه می‌تواند به‌طور ضمنی اثر آن‌ها را از داده‌ها بیاموزد. فقدان بررسی صریح این فرض، یکی از خلأهای اساسی در ادبیات این حوزه محسوب می‌شود.

بر این اساس، می‌توان نتیجه گرفت که تاکنون مطالعه‌ای جامع که به‌طور هم‌زمان به مسئلهٔ تخمین عمق متریک، پایداری نسبت به تغییر پارامترهای دوربین و ارزیابی مبتنی بر معیارهای حساس به مقیاس فیزیکی بپردازد، به‌صورت منسجم ارائه نشده است. پژوهش حاضر با تمرکز بر این خلأ، تلاش می‌کند گامی در جهت پر کردن این شکاف برداشته و چارچوبی برای تحلیل و بهبود پایداری عمق متریک در تخمین عمق تک‌دوربینه ارائه دهد.

۴.۱ اهداف و سوالات تحقیق

با توجه به خلأهای شناسایی‌شده در ادبیات موضوع و چالش‌های موجود در تخمین عمق تک‌دوربینه، پژوهش حاضر با هدف بررسی امکان دستیابی به تخمین عمق متریک پایدار در چارچوب یادگیری عمیق انجام شده است. تمرکز اصلی این تحقیق بر تحلیل نقش هندسهٔ دوربین در فرآیند تخمین عمق و ارزیابی تأثیر لحاظ‌کردن اطلاعات

[1] Non-metric Metrics

درونی دوربین بر پایداری و دقت خروجی مدل‌ها است.

هدف کلی این پژوهش، بررسی این مسئله است که آیا می‌توان با استفاده از اطلاعات صریح مربوط به مشخصات درونی دوربین، محدودیت‌های ذاتی تخمین عمق غیرمتریک را کاهش داد و به تخمینی دست یافت که نسبت به تغییر شرایط تصویربرداری پایدارتر باشد. در راستای این هدف کلی، اهداف جزئی زیر دنبال می‌شوند:

- بررسی میزان وابستگی مدل‌های تخمین عمق تک‌دوربینه به نوع دیتاست و پارامترهای دوربین؛
- تحلیل تفاوت رفتار معیارهای ارزیابی غیرمتریک و متریک در مواجهه با تغییرات شرایط تصویربرداری؛
- ارزیابی امکان بهبود دقت و پایداری تخمین عمق متریک با استفاده از اطلاعات هندسی صریح.

بر این اساس، سوالات اصلی تحقیق به‌صورت زیر مطرح می‌شوند:

۱. آیا مدل‌های رایج تخمین عمق تک‌دوربینه در برابر تغییر دیتاست و پارامترهای درونی دوربین رفتار پایداری از خود نشان می‌دهند؟

۲. آیا استفاده از معیارهای غیرمتریک می‌تواند ناپایداری‌های مرتبط با مقیاس فیزیکی عمق را پنهان کند؟

۳. آیا لحاظ‌کردن اطلاعات درونی دوربین می‌تواند به بهبود تخمین عمق متریک و کاهش حساسیت مدل نسبت به تغییرات دوربین منجر شود؟

جدول ۱.۱ نگاشت این سوالات را به فرضیه‌های تجربی متناظر و به بخشی از پایان‌نامه که شاهد آن ارائه می‌شود نشان می‌دهد. این نگاشت، مسیر خواندن فصل چهارم را نیز مشخص می‌کند.

پاسخ به این سوالات، چارچوب اصلی تحلیل‌های تجربی ارائه‌شده در فصل چهارم را تشکیل می‌دهد و مبنای نتیجه‌گیری‌های نهایی پژوهش حاضر قرار می‌گیرد.

۵.۱ ساختار پایان‌نامه

این پایان‌نامه در پنج فصل سازمان‌دهی شده است که هر فصل به بخش مشخصی از فرآیند پژوهش اختصاص دارد.

در فصل اول، مقدمه‌ای بر مسئلهٔ تخمین عمق تک‌دوربینه ارائه شد و با تبیین اهمیت موضوع، چالش‌های موجود و خلأهای پژوهشی، اهداف و سوالات تحقیق مطرح گردید.

جدول ۱.۱: نگاشت سوالات تحقیق به فرضیه‌های تجربی متناظر و بخش‌هایی که شاهد تجربی هر یک در آن‌ها ارائه می‌شود.

ردیف	سوال تحقیق	فرضیهٔ تجربی متناظر	شاهد اصلی در این پایان‌نامه
۱	پایداری مدل‌های رایج در برابر تغییر دیتاست و پارامترهای درونی دوربین	فرضیهٔ اول: مدل پایهٔ Depth Anything V2 در معیارهای متریک به دیتاست و دوربین حساس است	جدول ۱.۳ و بخش ۱.۴.۴
۲	توانایی معیارهای غیرمتریک در پنهان‌کردن ناپایداری مقیاس فیزیکی	فرضیهٔ سوم: بهبود متریک بدون افت ساختار نسبی به دست می‌آید و پروتکل نسبی نسبت به مقیاس نابیناست	بخش‌های ۵.۵.۴ و ۴.۹.۴
۳	اثر لحاظ صریح ماتریس مشخصهٔ دوربین بر دقت و پایداری عمق متریک	فرضیهٔ دوم: ورودی صریح K خطای متریک را به‌صورت معنادار کاهش می‌دهد	بخش‌های ۶.۵.۴، ۷.۵.۴ و ۸.۵.۴
۴	اثبات اینکه بهبود مشاهده‌شده واقعاً از مصرف K می‌آید، نه هم‌بستگی جانبی	شرط لازم برای اعتبار فرضیهٔ دوم: پاسخ خروجی به K باید کمی و مطابق هندسهٔ روزنه‌ای باشد	آزمون جاروی K (۱.۶.۴) و آزمون حذف K (۲.۶.۴)

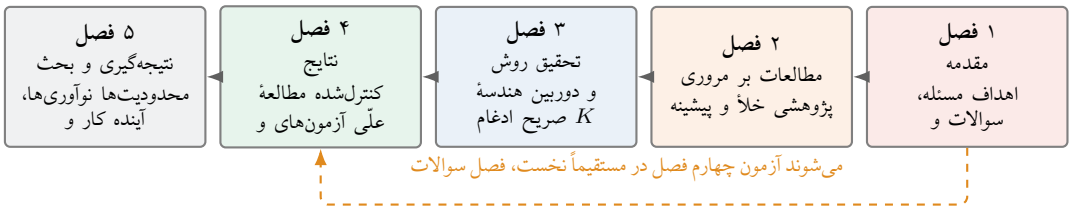
فصل دوم به مرور ادبیات موضوع اختصاص دارد. در این فصل، پژوهش‌های پیشین در حوزهٔ تخمین عمق تک‌دوربینه، روش‌های متریک و غیرمتریک، نقش هندسهٔ دوربین و معیارهای ارزیابی مورد بررسی و تحلیل قرار می‌گیرند و جایگاه پژوهش حاضر نسبت به کارهای موجود مشخص می‌شود.

در فصل سوم، روش پیشنهادی تحقیق به‌صورت جامع تشریح می‌شود. این فصل شامل صورت‌بندی مسئله، تحلیل هندسی فرآیند تصویربرداری، معرفی معماری پایه، انگیزه و چارچوب روش پیشنهادی، راهبرد آموزش، توابع هزینه و جزئیات پیاده‌سازی است.

فصل چهارم به ارائهٔ تنظیمات آزمایشی، نتایج کمی و کیفی و تحلیل آن‌ها اختصاص دارد. در این فصل، عملکرد مدل‌ها بر روی دیتاست‌های مختلف بررسی شده و پایداری تخمین عمق نسبت به تغییر پارامترهای دوربین به‌صورت تجربی تحلیل می‌شود.

شکل ۲.۱ این ساختار را به‌صورت یک نمای کلی نشان می‌دهد.

در نهایت، فصل پنجم به بحث و نتیجه‌گیری اختصاص دارد. در این فصل، یافته‌های اصلی پژوهش جمع‌بندی شده، نوآوری‌های نظری و عملی تحقیق بیان می‌شود و پیشنهادهایی برای پژوهش‌های آتی ارائه می‌گردد.



شکل ۲.۱: ساختار پایان‌نامه و رابطهٔ منطقی میان فصل‌ها. سوالات و فرضیه‌های مطرح‌شده در فصل نخست (جدول ۱.۱) در فصل چهارم به‌صورت تجربی آزمون می‌شوند و فصل پنجم برداشت نهایی را ارائه می‌کند.

فصل ۲

مروری بر مطالعات انجام شده

۱.۲ مقدمه

تخمین عمق از تصاویر دوبعدی یکی از مسائل بنیادی در بینایی ماشین[1] و ادراک صحنه محسوب می‌شود که کاربردهای گسترده‌ای در حوزه‌هایی نظیر رباتیک، رانندگی خودکار[2]، واقعیت افزوده[3] و بازسازی سه‌بعدی[4] دارد [۱، ۳]. در میان شاخه‌های مختلف این مسئله، تخمین عمق تک‌دوربینه[5] به‌دلیل نیاز حداقلی به تجهیزات تصویربرداری، بیشترین توجه را به خود جلب کرده است.

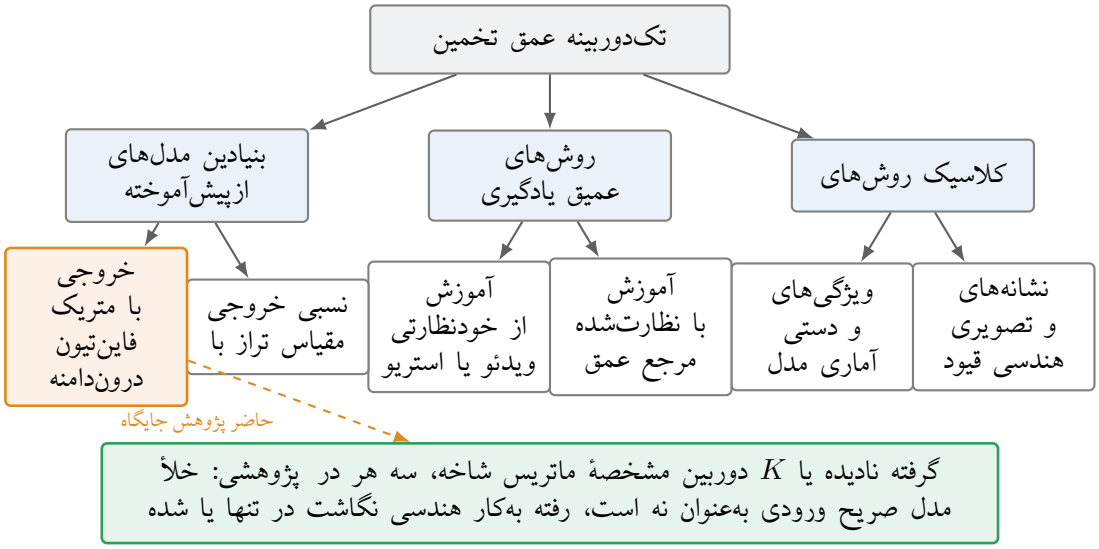
با این حال، تخمین عمق از یک تصویر منفرد ذاتاً یک مسئلهٔ ill-posed[6] محسوب می‌شود، زیرا نگاشت از فضای دوبعدی تصویر به فضای سه‌بعدی صحنه فاقد یکتایی ذاتی است [۴]. به‌همین دلیل، پژوهش‌های پیشین همواره تلاش کرده‌اند با اعمال قیود هندسی، آماری یا یادگیری از داده، این ابهام را کاهش دهند.

هدف از این فصل، ارائهٔ مروری ساختارمند بر پژوهش‌های انجام‌شده در حوزهٔ تخمین عمق تک‌دوربینه است. بدین منظور، ابتدا رویکردهای کلاسیک و سپس روش‌های مبتنی بر یادگیری عمیق مورد بررسی قرار می‌گیرند. در ادامه، تفاوت میان تخمین عمق متریک و غیرمتریک و نقش هندسهٔ دوربین در این مسئله مرور می‌شود تا جایگاه پژوهش حاضر نسبت به ادبیات موجود به‌روشنی مشخص گردد.

شکل ۱.۲ نمای کلی این دسته‌بندی و جایگاه پژوهش حاضر نسبت به آن را نشان می‌دهد.

---

[1] Computer Vision
[2] Autonomous Driving
[3] Augmented Reality
[4] 3D Reconstruction
[5] Monocular Depth Estimation
[6] بدحال



شکل ۱.۲: دسته‌بندی روش‌های تخمین عمق تک‌دوربینه و جایگاه پژوهش حاضر. سه شاخهٔ اصلی — روش‌های کلاسیک، روش‌های یادگیری عمیق و مدل‌های بنیادین ازپیش‌آموخته — در بخش‌های بعدی این فصل مرور می‌شوند. نکتهٔ مشترک هر سه شاخه آن است که پارامترهای درونی دوربین به‌صورت ورودی صریح مدل در نظر گرفته نمی‌شوند؛ همین نکته، خلأ پژوهشی مورد هدف این پایان‌نامه است.

۲.۲ روش‌های کلاسیک تخمین عمق تک‌دوربینه

پیش از گسترش روش‌های مبتنی بر یادگیری عمیق، تخمین عمق تک‌دوربینه عمدتاً بر پایهٔ روش‌های کلاسیک و استفاده از قیود هندسی و آماری انجام می‌شد [۱۰]. این روش‌ها معمولاً از نشانه‌های تصویری[1] مانند پرسپکتیو، انسداد، اندازهٔ نسبی اشیا، و تغییرات بافت برای استنتاج عمق استفاده می‌کردند.

برخی از رویکردها تلاش می‌کردند با مدل‌سازی صریح صحنه و اعمال فرض‌هایی دربارهٔ ساختار محیط، عمق نسبی نواحی مختلف تصویر را تخمین بزنند [۱]. در این روش‌ها، اطلاعاتی مانند خطوط همگرا، سطوح تخت، یا موقعیت افق نقش کلیدی ایفا می‌کردند. با وجود کارایی محدود در سناریوهای خاص، این روش‌ها به‌شدت به مفروضات اولیه وابسته بودند و قابلیت تعمیم پایینی داشتند.

دستهٔ دیگری از روش‌های کلاسیک بر یادگیری مبتنی بر ویژگی‌های دستی[2] تکیه داشتند. در این رویکردها، ابتدا ویژگی‌هایی از تصویر استخراج می‌شد و سپس با استفاده از مدل‌های آماری، نگاشتی بین این ویژگی‌ها و عمق تخمینی برقرار می‌گردید. اگرچه این روش‌ها نسبت به رویکردهای کاملاً هندسی انعطاف‌پذیرتر بودند، اما همچنان در مواجهه با صحنه‌های پیچیده و متنوع با محدودیت‌های جدی روبه‌رو بودند.

به‌طور کلی، روش‌های کلاسیک تخمین عمق تک‌دوربینه به‌دلیل وابستگی شدید به فرضیات صحنه و ناتوانی

[1] Visual Cues
[2] Hand-crafted Features

در بهره‌گیری مؤثر از داده‌های بزرگ، نتوانستند پاسخ‌گوی نیازهای کاربردهای مدرن باشند. این محدودیت‌ها زمینه‌ساز ظهور روش‌های مبتنی بر یادگیری عمیق در این حوزه شد.

۳.۲ روش‌های مبتنی بر یادگیری عمیق

با پیشرفت شبکه‌های عصبی عمیق[1] و در دسترس قرار گرفتن دیتاست‌های بزرگ، روش‌های مبتنی بر یادگیری عمیق به رویکرد غالب در تخمین عمق تک‌دوربینه تبدیل شدند [۲].

بخش قابل توجهی از این پژوهش‌ها از چارچوب‌های نظارت‌شده[2] استفاده می‌کنند [۵]. مدل با بهره‌گیری از نقشه‌های عمق مرجع آموزش داده می‌شود. این رویکردها معمولاً دقت بالایی در دیتاست‌های خاص ارائه می‌دهند، اما به‌شدت به کیفیت و دامنهٔ داده‌های آموزشی وابسته هستند و در سناریوهای خارج از توزیع آموزش، با افت عملکرد مواجه می‌شوند.

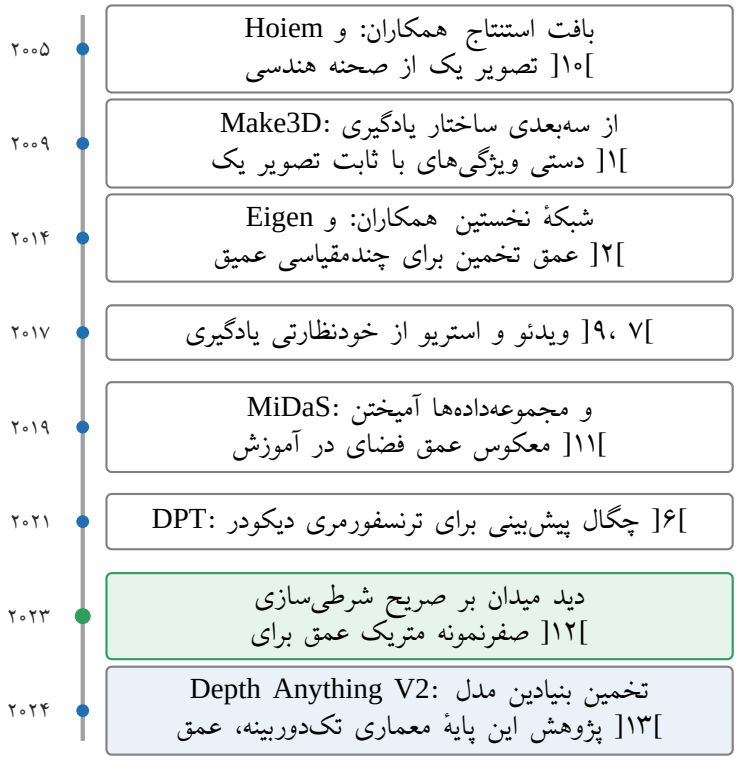
در مقابل، روش‌های بدون نظارت[3] تلاش می‌کنند با استفاده از قیود بازسازی تصویری و روابط هندسی ضمنی، مدل را بدون نیاز به داده‌های عمق مرجع آموزش دهند. این دسته از روش‌ها به‌ویژه در تخمین عمق تک‌دوربینه مورد توجه قرار گرفته‌اند، زیرا امکان آموزش مدل‌ها بر روی داده‌های واقعی و متنوع را فراهم می‌کنند.

با وجود موفقیت‌های قابل توجه، بیشتر روش‌های مبتنی بر یادگیری عمیق بر تخمین عمق نسبی تمرکز دارند و خروجی آن‌ها فاقد مقیاس فیزیکی مشخص است. در بسیاری از این مدل‌ها، فرض می‌شود که شبکه می‌تواند به‌طور ضمنی اطلاعات مربوط به هندسهٔ تصویربرداری را از داده‌ها بیاموزد. این فرض، اگرچه در برخی معیارهای ارزیابی پذیرفتنی به نظر می‌رسد، اما در کاربردهای نیازمند عمق متریک با چالش‌های جدی همراه است.

شکل ۲.۲ سیر تحول این حوزه را از روش‌های کلاسیک تا مدل‌های بنیادین امروزی و ظهور رویکردهای آگاه از هندسهٔ دوربین نشان می‌دهد.

شکل ۳.۲ این سه پارادایم را در کنار هم قرار می‌دهد و نشان می‌دهد که سیگنال آموزشی هر یک از کجا می‌آید و ماتریس مشخصهٔ دوربین در کدام مرحله — و آیا اصلاً — وارد فرآیند می‌شود.

[1] Deep Neural Networks
[2] Supervised Learning
[3] Self-supervised / Unsupervised



شکل ۲.۲: سیر تحول روش‌های تخمین عمق تک‌دوربینه. مسیر کلی از قیود هندسی دست‌ساز به سمت مدل‌های داده‌محور بزرگ حرکت کرده است؛ اما بازگشت اطلاعات هندسی صریح دوربین به فرآیند یادگیری تنها در سال‌های اخیر و به‌صورت محدود آغاز شده است.

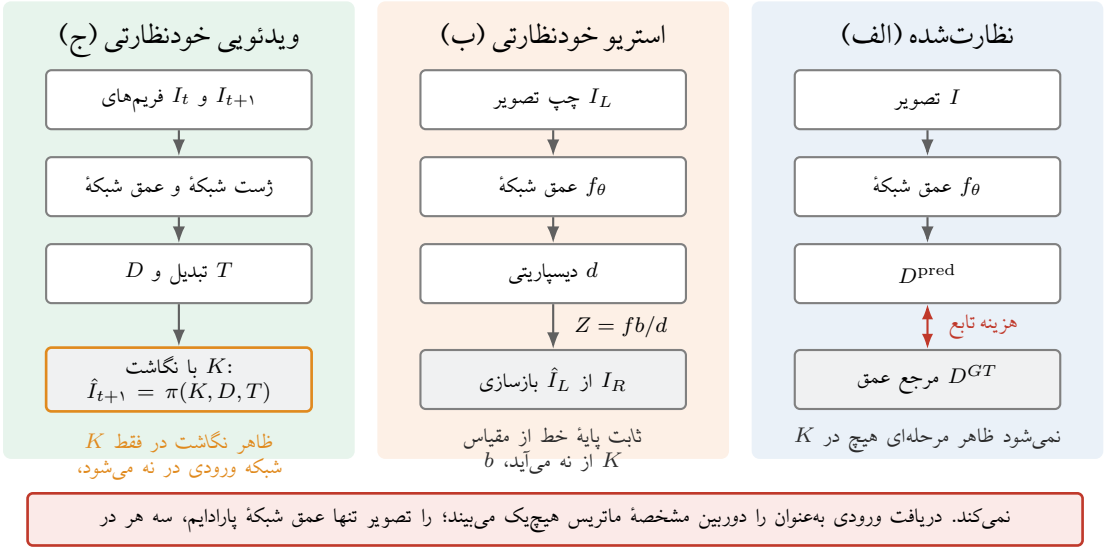
۴.۲ تخمین عمق متریک و غیرمتریک

یکی از تمایزهای اساسی در تخمین عمق تک‌دوربینه، تفاوت میان عمق غیرمتریک[1] و عمق متریک[2] است. در تخمین عمق غیرمتریک، مدل تنها قادر به بازیابی روابط نسبی میان نواحی مختلف صحنه است و مقیاس فیزیکی واقعی عمق به‌صورت ذاتی مبهم باقی می‌ماند. این ابهام مقیاس یکی از ویژگی‌های بنیادین مسئلهٔ تخمین عمق تک‌دوربینه محسوب می‌شود.

بخش قابل توجهی از روش‌های اولیهٔ مبتنی بر یادگیری عمیق، از جمله کارهای پیشگامانهٔ Eigen و همکاران [۲] و Eigen و Fergus [۵] بر تخمین عمق نسبی تمرکز داشتند. در این روش‌ها، اگرچه ساختار کلی صحنه با دقت مناسبی بازسازی می‌شد، اما خروجی مدل فاقد مقیاس فیزیکی پایدار بود و برای کاربردهای متریک نیازمند نرمال‌سازی یا کالیبراسیون پس‌پردازشی بود.

در ادامه، برخی پژوهش‌ها تلاش کردند با استفاده از قیود هندسی یا داده‌های اضافی، تخمین عمق متریک را

[1] Non-metric Depth
[2] Metric Depth



شکل ۳.۲: سه پارادایم نظارتی رایج در تخمین عمق تک‌دوربینه و جایگاه ماتریس مشخصهٔ دوربین در هر یک. (الف) در آموزش نظارت‌شده [۲]، سیگنال آموزشی از عمق مرجع می‌آید و K اصلاً وارد محاسبات نمی‌شود؛ مقیاس آموخته‌شده در عمل مقیاس دوربینِ دادهٔ آموزش است. (ب) در خودنظارتی استریو [۹]، مقیاس از خط پایهٔ ثابت دوربین‌های استریو استخراج می‌شود و بنابراین به تنظیم خاص آن مجموعه‌داده گره می‌خورد. (ج) در خودنظارتی ویدئویی [۷]، ماتریس K برای نگاشت میان نماها استفاده می‌شود، اما شبکهٔ عمق آن را نمی‌بیند و نمی‌تواند خروجی خود را بر اساس آن تنظیم کند. تمایز میان «استفاده از K در تابع هزینه» و «دادن K به شبکه» — که در شکل با رنگ متفاوت مشخص شده است — نقطهٔ آغاز روش پیشنهادی این پژوهش است.

ممکن سازند. برای مثال، روش‌های مبتنی بر یادگیری خودنظارتی که از بازسازی دیدگاه[1] استفاده می‌کنند، مانند کار Godard و همکاران [۹]، توانستند بدون استفاده از دادهٔ عمق مرجع، مدل‌هایی برای تخمین عمق آموزش دهند. با این حال، این روش‌ها نیز به‌طور معمول تنها به عمق نسبی دست می‌یابند و مقیاس خروجی آن‌ها وابسته به شرایط تصویربرداری و دادهٔ آموزشی است.

برخی مطالعات نشان داده‌اند که مدل‌های آموزش‌دیده با معیارهای غیرمتریک ممکن است در معیارهای نرمال‌شده عملکرد مطلوبی داشته باشند، اما در معیارهای حساس به مقیاس فیزیکی، مانند خطای جذر میانگین مربعات[2] افت قابل توجهی از خود نشان دهند [۶]. این مسئله به‌ویژه در هنگام تغییر نوع دوربین یا پارامترهای تصویربرداری بیشتر نمایان می‌شود.

در مجموع، می‌توان گفت که اگرچه تخمین عمق غیرمتریک برای بسیاری از کاربردهای ادراکی کافی است، اما برای کاربردهایی که نیازمند مقیاس واقعی هستند، مانند رباتیک و رانندگی خودکار، تخمین عمق متریک پایدار همچنان یک چالش حل‌نشده باقی مانده است.

[1] View Synthesis
[2] Root Mean Squared Error

۵.۲ نقش هندسهٔ دوربین و اطلاعات درونی

هندسهٔ تصویربرداری نقش بنیادینی در نگاشت فضای سه‌بعدی صحنه به تصویر دوبعدی ایفا می‌کند. مدل‌های کلاسیک تصویربرداری، مانند مدل دوربین سوراخ‌سوزنی، به‌روشنی نشان می‌دهند که پارامترهای درونی دوربین[1] نظیر فاصلهٔ کانونی و مرکز تصویر، به‌طور مستقیم بر رابطهٔ بین مختصات تصویری و عمق فیزیکی تأثیر می‌گذارند [۳].

با وجود این واقعیت هندسی، بخش عمده‌ای از روش‌های مبتنی بر یادگیری عمیق اطلاعات درونی دوربین را یا به‌طور کامل نادیده می‌گیرند و یا فرض می‌کنند که شبکه می‌تواند اثر آن‌ها را به‌صورت ضمنی از داده‌ها بیاموزد. این فرض به‌ویژه در روش‌هایی که بر داده‌های متنوع با دوربین‌های متفاوت آموزش داده می‌شوند، چالش‌برانگیز است.

برخی مطالعات نشان داده‌اند که تغییر پارامترهای درونی دوربین، به‌ویژه فاصلهٔ کانونی، می‌تواند منجر به تغییرات معنادار در مقیاس عمق تخمینی شود، حتی زمانی که صحنهٔ تصویربرداری ثابت باقی می‌ماند [۷]. این پدیده بیانگر آن است که نادیده‌گرفتن اطلاعات هندسی صریح می‌تواند به ناپایداری خروجی مدل‌ها منجر شود.

در سال‌های اخیر، تلاش‌هایی برای ترکیب صریح دانش هندسی با مدل‌های یادگیری عمیق صورت گرفته است. برای نمونه، برخی روش‌ها از پارامترهای کالیبراسیون در مراحل آموزش یا ارزیابی استفاده کرده‌اند [۸]. با این حال، استفادهٔ نظام‌مند و صریح از ماتریس مشخصهٔ دوربین[2] به‌عنوان ورودی مدل تخمین عمق هنوز به‌طور گسترده مورد بررسی قرار نگرفته است.

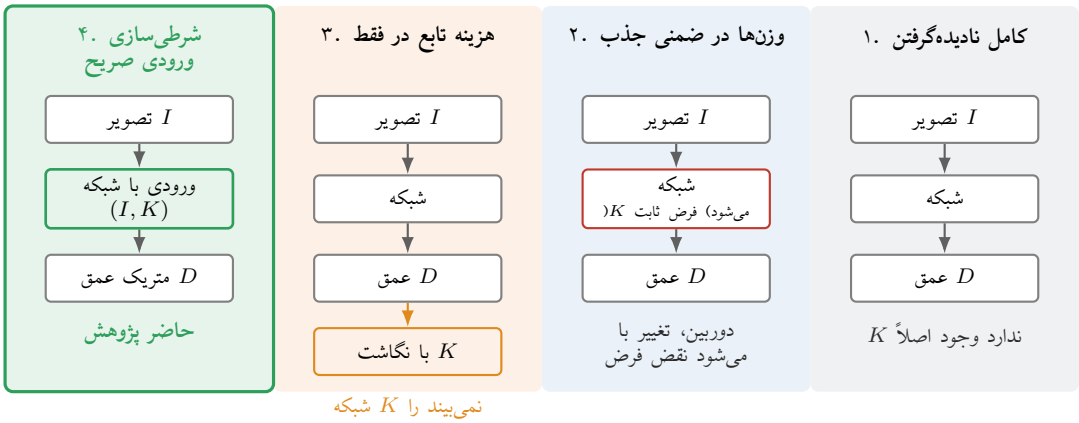
شکل ۴.۲ چهار راهبرد ممکن برای برخورد با ماتریس مشخصهٔ دوربین را در کنار هم قرار می‌دهد. سه راهبرد نخست، تقریباً تمام ادبیات موجود را پوشش می‌دهند و راهبرد چهارم، موضوع پژوهش حاضر است.

این مشاهدات نشان می‌دهد که نادیده‌گرفتن هندسهٔ دوربین یکی از عوامل اصلی ناپایداری عمق متریک در روش‌های موجود است. از این‌رو، تحلیل نقش اطلاعات درونی دوربین و بررسی تأثیر آن بر پایداری تخمین عمق می‌تواند مسیر روشنی برای بهبود مدل‌های تخمین عمق تک‌دوربینه فراهم کند.

۶.۲ جمع‌بندی فصل و جایگاه پژوهش حاضر

مرور ارائه‌شده در این فصل را می‌توان در قالب یک مقایسهٔ ساختاریافته میان کارهای شاخص این حوزه خلاصه کرد. جدول ۱.۲ نمونه‌ای از مهم‌ترین رویکردهای مرورشده را از چهار منظر مقایسه می‌کند: نوع سیگنال

[1] Camera Intrinsics
[2] Camera Intrinsic Matrix



شکل ۴.۲: چهار راهبرد ممکن برای برخورد با ماتریس مشخصهٔ دوربین در یک مدل تخمین عمق. راهبردهای ۱ و ۲ در روش‌های نظارت‌شده و بیشتر مدل‌های بنیادین رایج‌اند و راهبرد ۳ در روش‌های خودنظارتی ویدئویی به‌کار می‌رود. نکتهٔ مشترک هر سه آن است که خود شبکه هرگز مقدار K را دریافت نمی‌کند و بنابراین نمی‌تواند خروجی خود را متناسب با هندسهٔ دوربینِ تصویر ورودی تنظیم کند. راهبرد ۴، که موضوع این پایان‌نامه است، K را به یک ورودی درجه‌یک مدل تبدیل می‌کند.

نظارتی، نوع خروجی عمق، نحوهٔ برخورد با پارامترهای درونی دوربین، و محدودیت کلیدی هر رویکرد از منظر پایداری مقیاس متریک.

بر پایهٔ این مقایسه، خلأ پژوهشی این حوزه را می‌توان در چهار محور صورت‌بندی کرد که جدول ۲.۲ آن‌ها را در کنار پاسخ پژوهش حاضر قرار می‌دهد.

در مجموع، مرور ادبیات نشان می‌دهد که پرسش «آیا ورود صریح ماتریس مشخصهٔ دوربین به مدل، عمق متریک پایدارتری تولید می‌کند؟» تاکنون به‌صورت یک آزمایش کنترل‌شده و با بازوی هم‌داده پاسخ داده نشده است. فصل سوم، چارچوب روش‌شناختی لازم برای طراحی چنین آزمایشی را ارائه می‌کند.

جدول ۱.۲: مقایسهٔ ساختاریافتهٔ رویکردهای شاخص تخمین عمق تک‌دوربینه از منظر نوع نظارت، نوع خروجی، نحوهٔ برخورد با ماتریس مشخصهٔ دوربین K و محدودیت کلیدی. ستون «برخورد با K» نشان می‌دهد که در بخش عمدهٔ ادبیات، K یا اصلاً حضور ندارد یا تنها در نگاشت هندسی تابع هزینه به‌کار می‌رود، نه به‌عنوان ورودی مدل.

رویکرد	سال	نوع نظارت	نوع خروجی	برخورد با K	محدودیت کلیدی
Hoiem و همکاران [۱۰]	۲۰۰۵	کلاسیک، بدون یادگیری عمیق	نسبی	ندارد	وابستگی شدید به فرض‌های ساختار صحنه
Make3D [۱]	۲۰۰۹	نظارت‌شده با ویژگی دستی	نسبی	ندارد	تعمیم ضعیف به صحنه‌های جدید
Eigen و همکاران [۲]	۲۰۱۴	نظارت‌شده	متریک درون‌دامنه	ندارد	مقیاس آموخته‌شده مختص دوربین دادهٔ آموزش است
Eigen و Fergus [۵]	۲۰۱۵	نظارت‌شدهٔ چندوظیفه‌ای	متریک درون‌دامنه	ندارد	ارزیابی عمدتاً با معیارهای نرمال‌شده
Godard و همکاران [۹]	۲۰۱۷	خودنظارتی استریو	نسبی	ضمنی، از راه خط پایهٔ ثابت	مقیاس وابسته به تنظیم استریوی آموزش
Zhou و همکاران [۷]	۲۰۱۷	خودنظارتی ویدئویی	نسبی	در نگاشت تابع هزینه، نه ورودی شبکه	ابهام مقیاس رفع‌نشده باقی می‌ماند
Monodepth2 [۱۴]	۲۰۱۹	خودنظارتی	نسبی	ضمنی	ارزیابی نیازمند تراز مقیاس است
MiDaS [۱۱]	۲۰۱۹	نظارت‌شدهٔ آمیخته	نسبی، در فضای عمق معکوس	ندارد	خروجی ذاتاً غیرمتریک است
PackNet [۱۵]	۲۰۲۰	خودنظارتی	نسبی، با مقیاس از سیگنال کمکی	غیرمستقیم	نیازمند سیگنال خارجی برای بازیابی مقیاس
DPT [۶]	۲۰۲۱	نظارت‌شدهٔ آمیخته	نسبی و متریک درون‌دامنه	ندارد	حساسیت به تغییر مشخصات دوربین
استریوی چندنمای آگاه از هندسه [۱۶]	۲۰۲۲	چندنما	متریک	صریح، در حجم هزینه	نیازمند چند نما؛ تک‌تصویری نیست
شرطی‌سازی بر میدان دید [۱۲]	۲۰۲۳	مدل انتشاری	متریک صفرنمونه	صریح، در سطح میدان دید	هزینهٔ محاسباتی بالای استنتاج
Depth Anything V2 [۱۳]	۲۰۲۴	تقطیر در مقیاس بزرگ	نسبی، به‌همراه شاخهٔ متریک	ندارد	مقیاس متریک به دامنهٔ جدید منتقل نمی‌شود
پژوهش حاضر	—	**فاین‌تیون نظارتی کنترل‌شده**	**متریک**	**ورودی صریح به انکودر**	**حساسیت به صحت کالیبراسیون**

جدول ۲.۲: صورت‌بندی خلأ پژوهشی شناسایی‌شده در ادبیات موضوع و پاسخ متناظر پژوهش حاضر به هر محور.

محور	وضعیت غالب در ادبیات	پاسخ پژوهش حاضر
بازنمایی هندسهٔ دوربین	K یا نادیده گرفته می‌شود یا فرض می‌شود شبکه اثر آن را ضمنی می‌آموزد	K به یک توکن نهان تبدیل و به‌صورت صریح به دنبالهٔ توکن‌های انکودر تزریق می‌شود
معیار سنجش	غلبهٔ معیارهای نرمال‌شده که نسبت به مقیاس مطلق نابینا هستند	گزارش هم‌زمان پروتکل نسبی و پروتکل متریک بدون تراز، و تحلیل تفاوت رفتار آن‌ها
طراحی آزمایش	مقایسه با مدل‌های منتشرشده، بدون بازوی کنترل هم‌داده	زوج‌سازی کامل مدل پیشنهادی و مدل کنترل با داده، ابرپارامتر و بذر تصادفی یکسان
اثبات علیت	اکتفا به هم‌بستگی میان افزودن اطلاعات هندسی و بهبود معیارها	آزمون‌های سازوکاری جاروی K و حذف و تحریف K برای سنجش مستقیم مصرف اطلاعات دوربین

فصل ۳

روش تحقیق

۱.۳ مقدمه

در این فصل، روش پیشنهادی برای شرطی‌سازی یک مدل بنیادین تخمین عمق تک‌چشمی با پارامترهای درونی دوربین تشریح می‌شود. تمرکز اصلی بر معماری، نحوهٔ ادغام اطلاعات هندسی و فرآیند واقعی آموزش و ارزیابی است. مدل پایهٔ Depth Anything V2 از ادبیات یادگیری خودنظارتی[1] و تقطیر بهره می‌گیرد، اما فاین‌تیون نهایی این پژوهش با عمق مرجع انجام شده است؛ ازاین‌رو، در سراسر فصل میان پیش‌آموزش مدل پایهٔ Depth Anything V2، طرح اختیاری تقطیر و آموزش نظارتی آزمایش‌های نهایی تمایز گذاشته می‌شود.

مدل پایهٔ Depth Anything V2 یک معماری encoder-decoder مبتنی بر Transformer Vision را برای تخمین عمق چگال از تصاویر تک‌چشمی ارائه می‌دهد [۱۳]. اگرچه این مدل در حالت پایه عملکرد قابل‌توجهی در سناریوهای مختلف نشان می‌دهد، اما همانند بسیاری از روش‌های موجود، اطلاعات درونی دوربین به‌صورت صریح در ورودی مدل لحاظ نمی‌شود [۲، ۹، ۶]. این موضوع می‌تواند منجر به ناپایداری تخمین عمق در مواجهه با تغییر پارامترهای دوربین، به‌ویژه در داده‌های چندمنبعه یا تنظیمات دنیای واقعی، شود.

بر این اساس، در روش پیشنهادی این پژوهش، ماتریس مشخصه دوربین[2] به‌عنوان یک ورودی اضافی و ساختاریافته به بخش encoder مدل افزوده شده است. هدف از این کار، فراهم‌کردن امکان استفاده مستقیم از اطلاعات هندسی دوربین در فرآیند یادگیری و تأثیرگذاری آن بر مکانیزم attention در لایه‌های Transformer است. بدین ترتیب، مدل قادر خواهد بود نگاشت تصویر به عمق را نه‌تنها بر اساس محتوای بصری صحنه، بلکه

[1] Self-supervised Learning
[2] Camera Intrinsic Matrix (K)

با درنظرگرفتن ویژگی‌های تصویربرداری دوربین انجام دهد.

در ادامه این فصل، ابتدا فرمول‌بندی دقیق مسئله و مدل هندسی دوربین ارائه می‌شود. سپس معماری Depth Anything V2 به‌صورت جزئی بررسی شده و تغییرات اعمال‌شده برای ادغام ماتریس مشخصه دوربین توضیح داده می‌شود. در بخش‌های پایانی، جزئیات مربوط به فرآیند آموزش، توابع هزینه مورد استفاده و ملاحظات پیاده‌سازی بیان خواهد شد.

۲.۳ صورت‌بندی مسئلهٔ تخمین عمق تک‌دوربینه

مسئلهٔ تخمین عمق تک‌دوربینه[1] به فرایند برآورد فاصلهٔ هر نقطه از صحنه نسبت به مرکز دوربین، تنها با استفاده از یک تصویر دوبعدی اطلاق می‌شود. در این مسئله، ورودی سیستم یک تصویر رنگی دوبعدی تعریف می‌شود و خروجی آن یک نگاشت عمق چگال است که به هر پیکسل تصویر، یک مقدار اسکالر متناظر با عمق نسبت داده می‌شود. با وجود ظاهر سادهٔ این تعریف، تخمین عمق تک‌دوربینه ذاتاً یک مسئلهٔ بدوضعیت محسوب می‌شود [۳، ۱] زیرا نگاشت از فضای دوبعدی تصویر به فضای سه‌بعدی صحنه، یکتا نبوده و چندین پیکربندی هندسی متفاوت می‌توانند تصویر یکسانی را ایجاد کنند.

در چارچوب یادگیری عمیق، این مسئله به‌صورت یادگیری یک تابع غیرخطی پارامتریزه‌شده با پارامترهای θ مدل‌سازی می‌شود که نگاشت زیر را تقریب می‌زند:

$$f_\theta : \mathbb{R}^{H\times W\times ۳} \to \mathbb{R}^{H\times W}, \tag{۱.۳}$$

که در آن، H و W به‌ترتیب ارتفاع و عرض تصویر ورودی بوده و خروجی، نگاشت عمق چگال متناظر با تصویر ورودی است. مقدار عمق پیش‌بینی‌شده در هر پیکسل، فاصلهٔ آن نقطه از مرکز اپتیکی دوربین را در راستای محور دید نمایش می‌دهد.

در روش‌های بدون نظارت یا خودنظارتی، برخلاف روش‌های نظارت‌شده، برچسب عمق صریح در دسترس نیست و یادگیری معمولاً بر قیود هندسی، سازگاری تصویری و بازسازی میان نماهای متوالی تکیه دارد [۷، ۹]. این پیشینه انگیزهٔ عنوان و صورت‌بندی هندسی پژوهش است. بااین‌حال، آزمایش کنترل‌شدهٔ نهایی برای اندازه‌گیری دقیق سهم ماتریس مشخصه، از عمق مرجع استفاده می‌کند و بنابراین باید یک مرحلهٔ فاین‌تیون نظارتی تلقی شود.

به‌صورت صوری، اگر I_t تصویر ورودی در زمان t باشد، هدف یادگیری تابعی است که نگاشت عمق D_t را

[1] Monocular Depth Estimation

به‌گونه‌ای تولید کند که با استفاده از آن و با در نظر گرفتن حرکت دوربین، تصویر I_{t+1} قابل بازسازی از I_t باشد. این بازسازی مبتنی بر هندسهٔ تصویربرداری و مدل دوربین بوده و به‌طور مستقیم به پارامترهای درونی دوربین وابسته است [۳]. با این حال، بخش قابل‌توجهی از روش‌های موجود، یا به‌صورت ضمنی این پارامترها را نادیده می‌گیرند یا فرض می‌کنند که تمامی تصاویر با یک دوربین ایده‌آل و یکسان ثبت شده‌اند.

یکی از تمایزهای اساسی در مسئلهٔ تخمین عمق تک‌دوربینه، تفاوت میان عمق نسبی و عمق مطلق است. بسیاری از روش‌های بدون نظارت قادر به پیش‌بینی عمق تنها تا یک ضریب مقیاس نامشخص هستند؛ بدین معنا که ساختار نسبی صحنه حفظ می‌شود، اما مقادیر عمق از نظر متریک معنا ندارند. این مسئله به‌ویژه در کاربردهایی که نیازمند دقت عددی در واحدهای فیزیکی هستند، نظیر ناوبری رباتیک یا بازسازی سه‌بعدی، یک محدودیت جدی محسوب می‌شود.

در مقابل، تخمین عمق مطلق یا متریک مستلزم آن است که نگاشت عمق خروجی، مستقیماً قابل تفسیر در واحدهای فیزیکی نظیر متر باشد. تحقق این هدف بدون در نظر گرفتن مشخصات هندسی دوربین عملاً امکان‌پذیر نیست، زیرا تبدیل میان مختصات تصویری و مختصات سه‌بعدی جهان، به‌طور مستقیم به پارامترهای درونی دوربین وابسته است. این وابستگی از طریق ماتریس مشخصهٔ دوربین مدل‌سازی می‌شود که نقش کلیدی در تعیین مقیاس و اعوجاج هندسی تصویر دارد.

در این پژوهش، مسئله به‌صورت زیر تعریف می‌شود: هدف، یادگیری تابعی است که علاوه بر تصویر ورودی، پارامترهای درونی دوربین را نیز دریافت کرده و نگاشت عمقی سازگار با هندسهٔ تصویربرداری تولید کند. این تعریف به نوع سیگنال نظارتی وابسته نیست و می‌تواند در آموزش نظارتی، خودنظارتی یا تقطیری به‌کار رود؛ پیاده‌سازی نهایی حاضر از حالت نظارتی استفاده می‌کند. به‌صورت نمادین:

$$f_\theta : (I, K) \to D, \tag{۲.۳}$$

که در آن، I تصویر ورودی، K ماتریس مشخصهٔ دوربین، و D نگاشت عمق خروجی است. در این صورت‌بندی، ماتریس مشخصه به‌عنوان بخشی از ورودی مدل در نظر گرفته می‌شود و مدل می‌آموزد که چگونه اطلاعات هندسی دوربین را در فرآیند استنتاج عمق لحاظ کند.

این بازفرمول‌بندی، امکان تحلیل صریح اثر تغییرات دوربین بر خروجی مدل را فراهم می‌سازد و زمینه را برای تبیین تفاوت رفتار معیارهای ارزیابی متریک و غیرمتریک مهیا می‌کند. به‌ویژه، معیارهایی نظیر خطای میانگین مربعات ریشه[1] که به مقیاس فیزیکی حساس هستند، مستقیماً از نحوهٔ مدل‌سازی پارامترهای دوربین تأثیر

[1] Root Mean Square Error

می‌پذیرند، در حالی که معیارهای نسبی‌تر نظیر خطای نسبی مطلق[1] کمتر به این تغییرات واکنش نشان می‌دهند. این تمایز، در فصول بعدی و به‌ویژه در تحلیل نتایج تجربی، نقش محوری ایفا خواهد کرد.

۳.۳ هندسهٔ دوربین و نقش پارامترهای درونی

مسئلهٔ تخمین عمق، به‌طور بنیادین ریشه در هندسهٔ تصویربرداری دارد و هرگونه نگاشت میان فضای دوبعدی تصویر و ساختار سه‌بعدی صحنه، ناگزیر وابسته به مدل دوربین مورد استفاده است. در تخمین عمق تک‌دوربینه، این وابستگی حتی پررنگ‌تر می‌شود، زیرا تمام اطلاعات مربوط به ساختار فضایی صحنه باید از روی یک تصویر و با اتکا به قیود هندسی استنتاج شود. از این رو، درک دقیق هندسهٔ دوربین و نقش پارامترهای درونی آن، برای تحلیل رفتار مدل‌های تخمین عمق، به‌ویژه در تنظیمات بدون نظارت، ضروری است.

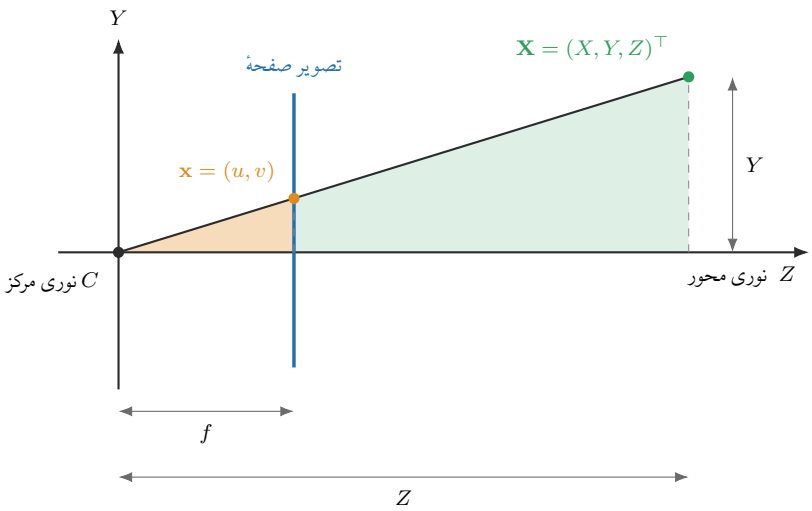
در این بخش، ابتدا مدل هندسی دوربین مورد استفاده معرفی می‌شود، سپس ماتریس مشخصهٔ دوربین به‌صورت صوری تعریف شده و در ادامه، تأثیر مستقیم پارامترهای درونی بر مقیاس و پایداری تخمین عمق مورد بررسی قرار می‌گیرد.

۱.۳.۳ مدل دوربین سوراخ‌سوزنی

در اغلب روش‌های بینایی ماشین، از مدل دوربین سوراخ‌سوزنی به‌عنوان یک تقریب هندسی استاندارد برای فرآیند تصویربرداری استفاده می‌شود [۳]. در این مدل، فرض می‌شود که تمام پرتوهای نوری از یک نقطهٔ واحد، موسوم به مرکز اپتیکی، عبور کرده و بر روی صفحهٔ تصویر نگاشت می‌شوند. اگرچه این مدل بسیاری از پدیده‌های فیزیکی نظیر اعوجاج لنز را نادیده می‌گیرد، اما برای تحلیل هندسی و فرمول‌بندی ریاضی مسئلهٔ تخمین عمق، کفایت لازم را دارد.

در چارچوب این مدل، یک نقطهٔ سه‌بعدی در مختصات جهان به‌صورت $\mathbf{X} = (X, Y, Z)^\top$ نمایش داده می‌شود. این نقطه پس از تبدیل به دستگاه مختصات دوربین و انجام عمل تصویرسازی، به یک نقطهٔ دوبعدی روی صفحهٔ تصویر با مختصات پیکسلی (u, v) نگاشت می‌شود. این نگاشت، تحت فرضیات مدل سوراخ‌سوزنی، یک نگاشت پرسپکتیو بوده و ذاتاً غیرخطی است.

[1] Absolute Relative Error



$$\boxed{\frac{u}{f} = \frac{X}{Z}, \qquad \frac{v}{f} = \frac{Y}{Z}}$$

شکل ۱.۳: مدل هندسی دوربین سوراخ‌سوزنی: نگاشت یک نقطهٔ سه‌بعدی $\mathbf{X}$ به نقطهٔ دوبعدی $\mathbf{x}$ روی صفحهٔ تصویر. دو مثلث هم‌رنگ‌شده متشابه‌اند و همین تشابه، رابطهٔ بنیادین $u/f = X/Z$ را نتیجه می‌دهد. این رابطه نشان می‌دهد که عمق Z و فاصلهٔ کانونی f تنها به‌صورت نسبت در تصویر ظاهر می‌شوند؛ بنابراین بدون دانستن f، مقدار Z از روی تصویر یکتا تعیین‌پذیر نیست (مقایسه کنید با شکل ۱.۱).

۲.۳.۳ ماتریس مشخصهٔ دوربین

رابطهٔ میان مختصات سه‌بعدی یک نقطه در دستگاه مختصات دوربین و مختصات دوبعدی آن روی تصویر، به‌وسیلهٔ ماتریس مشخصهٔ دوربین مدل‌سازی می‌شود. این ماتریس که معمولاً با نماد K نمایش داده می‌شود، پارامترهای درونی دوربین را در بر می‌گیرد و به‌صورت زیر تعریف می‌شود:

$$K = \begin{bmatrix} f_x & ۰ & c_x \\ ۰ & f_y & c_y \\ ۰ & ۰ & ۱ \end{bmatrix}, \tag{۳.۳}$$

که در آن، f_x و f_y فاصلهٔ کانونی مؤثر دوربین در راستاهای افقی و عمودی بر حسب پیکسل هستند و (c_x, c_y) مختصات نقطهٔ اصلی تصویر را مشخص می‌کنند. این پارامترها به ویژگی‌های فیزیکی دوربین، نظیر فاصلهٔ کانونی لنز، اندازهٔ سنسور، و رزولوشن تصویر وابسته‌اند.

با استفاده از این ماتریس، رابطهٔ میان یک نقطهٔ سه‌بعدی $\mathbf{X}$ و تصویر دوبعدی آن را می‌توان به‌شکل زیر نوشت:

$$\lambda \begin{bmatrix} u \\ v \\ ۱ \end{bmatrix} = K \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{۴.۳}$$

که در آن، λ یک ضریب مقیاس است که به عمق نقطه، یعنی مؤلفهٔ Z، وابسته است. این رابطه به‌وضوح نشان می‌دهد که مختصات تصویری هر پیکسل، نه‌تنها به موقعیت سه‌بعدی نقطه، بلکه به پارامترهای درونی دوربین نیز وابسته است [۳].

۳.۳.۳ نقش پارامترهای درونی در تخمین عمق

در مسئلهٔ تخمین عمق، هدف اصلی برآورد مؤلفهٔ Z برای هر پیکسل تصویر است. با توجه به رابطهٔ هندسی فوق، روشن است که تبدیل میان مختصات تصویری و عمق واقعی صحنه، بدون اطلاع از ماتریس مشخصهٔ دوربین، به‌طور کامل قابل تعیین نیست. در غیاب این اطلاعات، مدل تنها قادر به یادگیری روابط نسبی میان پیکسل‌ها خواهد بود و عمق به‌دست‌آمده فاقد مقیاس متریک مشخص خواهد بود.

این مسئله در روش‌های بدون نظارت، که یادگیری بر پایهٔ بازسازی تصویری انجام می‌شود، اهمیت دوچندان می‌یابد. در این روش‌ها، بازسازی یک تصویر از روی تصویر دیگر مستلزم انجام عملیات پروجکشن معکوس و نگاشت نقاط از یک نما به نمای دیگر است. این عملیات به‌طور مستقیم از ماتریس مشخصهٔ دوربین استفاده می‌کند و هرگونه خطا یا فرض نادرست دربارهٔ پارامترهای درونی، منجر به بازسازی نادرست و در نتیجه یادگیری عمق ناپایدار می‌شود.

در بسیاری از روش‌های پیشین، یا فرض شده است که تمام تصاویر با یک دوربین ثابت و با پارامترهای درونی یکسان ثبت شده‌اند، یا اطلاعات مربوط به دوربین به‌صورت ضمنی در وزن‌های شبکه جذب شده است. این فرضیات، در عمل و به‌ویژه در تنظیمات چنددوربینه یا داده‌های جمع‌آوری‌شده در شرایط واقعی، معتبر نیستند. تغییر در فاصلهٔ کانونی یا نقطهٔ اصلی تصویر می‌تواند نگاشت میان پیکسل و عمق را به‌طور معناداری تغییر دهد، حتی اگر محتوای صحنه ثابت باقی بماند.

پیامد مستقیم این موضوع، ناپایداری تخمین عمق متریک در مواجهه با داده‌های ناهمگن از نظر مشخصات دوربین است. معیارهایی که به مقیاس فیزیکی عمق حساس هستند، نظیر خطای میانگین مربعات ریشه، بیش از سایر معیارها تحت تأثیر این ناپایداری قرار می‌گیرند. در مقابل، معیارهای مبتنی بر روابط نسبی میان پیکسل‌ها،

که به مقیاس مطلق وابسته نیستند، رفتار پایدارتری از خود نشان می‌دهند.

بر این اساس، لحاظ صریح پارامترهای درونی دوربین در فرآیند تخمین عمق، نه‌تنها از منظر هندسی موجه است، بلکه برای دستیابی به عمق متریک پایدار و قابل تعمیم، امری ضروری به نظر می‌رسد. این مشاهده، مبنای نظری روش پیشنهادی این پژوهش را تشکیل می‌دهد که در آن، ماتریس مشخصهٔ دوربین به‌عنوان بخشی از ورودی مدل در نظر گرفته شده و نقش آن در فرآیند استنتاج عمق به‌صورت ساخت‌یافته مدل‌سازی می‌شود.

۴.۳ معماری پایه: مدل Depth Anything V2

مدل Depth Anything V2 یکی از جدیدترین چارچوب‌های ارائه‌شده برای تخمین عمق تک‌دوربینه یادگیری خودنظارتی مبتنی بر تقطیر است [۱۳] و با بهره‌گیری از معماری‌های مبتنی بر ترنسفورمر و به‌کارگیری راهبردهای بازنمایی چندمقیاسی، قادر است نگاشت‌های عمق باکیفیت و سازگار با جزئیات ساختاری تصویر تولید کند. در این مدل، ساختار اصلی شبکه بر مبنای یک معماری encoder-decoder بنا شده است که در آن، بخش encoder مسئول استخراج بازنمایی‌های سطح‌بالا و چندمقیاسی از تصویر ورودی بوده و بخش decoder وظیفهٔ بازگردانی این بازنمایی‌ها به یک نگاشت عمق چگال را بر عهده دارد. این معماری، به‌ویژه در نسخهٔ دوم مدل، با بهبودهای ساختاری مهمی همراه شده است که موجب افزایش دقت تخمین عمق در هر دو محیط درونی و بیرونی و نیز بهبود رفتار مدل در داده‌های ناهمگن شده است.

۱.۴.۳ ساختار کلی شبکهٔ encoder-decoder

در معماری Depth Anything V2، تصویر ورودی پس از اعمال پیش‌پردازش مناسب، به بخش encoder تغذیه می‌شود. خروجی این بخش مجموعه‌ای از بازنمایی‌های نهان در سطوح مختلف است که در ادامه توسط بخش decoder برای تولید نقشهٔ عمق استفاده می‌شوند. این ساختار دو مرحله‌ای به مدل اجازه می‌دهد تا از یک سو، توصیف‌های سطح‌بالای مبتنی بر محتوای معنایی تصویر را استخراج کند و از سوی دیگر، با بهره‌گیری از ساختارهای چندمقیاسی، جزئیات مکانی لازم برای تولید عمق دقیق را حفظ نماید. در این معماری، جریان اطلاعات میان encoder و decoder از طریق مسیرهای میان‌بُری برقرار است تا ویژگی‌های مکانی که ممکن است در تغییرمقیاس‌های مکرر از دست بروند، در فرآیند بازسازی نقشهٔ عمق حفظ شوند.

۲.۴.۳ بخش encoder مبتنی بر Transformer Vision

بخش encoder در این مدل بر پایهٔ معماری Transformer Vision طراحی شده است [۱۷] که در سال‌های اخیر به‌عنوان یکی از مؤثرترین ساختارها در بازنمایی‌های بصری مطرح شده است. در این بخش، تصویر ورودی به مجموعه‌ای از قطعه‌های منظم تقسیم شده و سپس هر قطعه به یک بردار نهان با ابعاد ثابت نگاشت می‌شود. این بردارها در ادامه وارد لایه‌های خودتوجهی چندسری[1] شده و وابستگی‌های غیرمحلی میان بخش‌های مختلف تصویر را مدل‌سازی می‌کنند. استفاده از سازوکار[2] توجه به مدل اجازه می‌دهد تا روابط بلندبرد میان اجزای تصویر به‌طور مستقیم در بازنمایی نهان لحاظ شود، امری که به‌ویژه در تخمین عمق تک‌دوربینه اهمیت اساسی دارد؛ زیرا رابطهٔ میان هندسهٔ صحنه و محتوای تصویری، اغلب به وابستگی‌های گستردهٔ مکانی متکی است.

در نسخهٔ دوم Depth Anything، بخش encoder علاوه بر بهبود در ساختار توجه و افزایش ظرفیت مدل، با به‌کارگیری دانش انتقالی از مدل‌های ازپیش‌آموزش‌دیدهٔ بزرگ، نظیر مدل‌های مبتنی بر داده‌های گستردهٔ طبقه‌بندی و بخش‌بندی، قادر شده است بازنمایی‌های بصری دقیق‌تری را تولید کند. این رویکرد موجب افزایش تعمیم‌پذیری مدل در صحنه‌های جدید و شرایط نوری متفاوت شده است.

۳.۴.۳ بخش decoder و راهبرد بازسازی نگاشت عمق

بخش decoder در Depth Anything V2 بر پایهٔ ساختاری از خانوادهٔ DPT طراحی شده است [۶]؛ ساختاری که به‌طور خاص برای بازسازی نگاشت‌های چگال، نظیر عمق یا نرمال سطح، توسعه یافته است. این بخش با ترکیب بازنمایی‌های چندمقیاسی استخراج‌شده از encoder و انجام فرآیندهای ادغام مکانی، نگاشت عمق نهایی را تولید می‌کند. در این فرآیند، لایه‌های افزایش تفکیک‌پذیری، همراه با سازوکارهای ادغام ویژگی‌ها، نقشی اساسی در بازگردانی جزئیات محلی ایفا می‌کنند. از آنجا که بخش encoder به‌سبب تغییرمقیاس‌های متعدد، بخشی از اطلاعات مکانی صحنه را از دست می‌دهد، وجود مسیرهای ادغام میان‌بُری که بازنمایی‌های سطح پایین‌تر را با سطوح بالاتر ترکیب می‌کنند، برای تولید عمق دقیق و دارای مرزبندی روشن ضروری است.

در نسخهٔ دوم این مدل، decoder به‌گونه‌ای بازطراحی شده است که بتواند به شیوه‌ای کارآمد با رزولوشن‌های بالاتر نیز سازگار شود. به‌طور خاص، شبکه قادر است ورودی‌هایی با ابعاد تا چند صد پیکسل در هر ضلع را پردازش کرده و در عین حال، پایداری خود را در تخمین عمق حفظ کند. این ویژگی به‌طور مستقیم از نیازهای کاربردهای بازسازی سه‌بعدی با وضوح بالا الهام گرفته است، کاربردهایی که در مطالعات اخیر، از جمله

[1] منظور از خودتوجهی چندسری همان Multi-Head Self-Attention است که در معماری ترنسفورمر معرفی شده و با استفاده از چندین سر توجه به‌صورت موازی، روابط متفاوتی میان ویژگی‌ها را مدل می‌کند.

[2] Attention

روش‌های افزایش وضوح تا حد 4K، اهمیت زیادی پیدا کرده‌اند.

در مجموع، ترکیب یک encoder مبتنی بر ترنسفورمر با یک decoder مبتنی بر ساختارهای بازسازی چگال، موجب شده است Depth Anything V2 بتواند میان بازنمایی سطح‌بالا و جزئیات مکانی تعادل مناسبی برقرار کند. این ویژگی، آن را به مدلی مناسب برای استفاده به‌عنوان معماری پایه در این پژوهش بدل می‌کند. در بخش‌های بعدی، نحوهٔ ادغام صریح ماتریس مشخصهٔ دوربین در این معماری، و نیز تأثیر آن بر کیفیت تخمین عمق متریک، به‌تفصیل بررسی می‌شود.

۵.۳ انگیزه‌ی صریح برای لحاظ پارامترهای درونی دوربین

در فرآیند تخمین عمق تک‌دوربینه[1]، رابطهٔ بنیادین میان تصویربرداری و هندسهٔ صحنه، مستقیما به چگونگی بازنمایی پارامترهای درونی دوربین[2] وابسته است. اغلب مطالعات پیشین، با فرض یکسان بودن مشخصات اپتیکی و هندسی در تمام تصاویر مجموعه‌داده، یا با اتکا به دانش ازپیش‌آموزش‌یافته مدل نسبت به ساختار معمول دوربین‌ها، از لحاظ صریح و صوری این پارامترها در چرخهٔ فرایند یادگیری صرف‌نظر کرده‌اند. [۹، ۷] این رویکرد، هرچند روی مجموعه‌های همگن (مانند دیتاست‌های کنترل‌شدهٔ آزمایشگاهی) می‌تواند به نتایجی قابل اتکاء منجر شود، اما به‌شدت نسبت به ناهمگونی‌های اپتیکی و تفاوت‌های ساختاری میان نمونه‌های تصویری حساس است؛ عاملی که در کاربردهای دنیای واقعی و نیز در انتقال مدل به مجموعه‌داده‌های ناهمگن، موجب ناپایداری و کاهش دقت تخمین عمق می‌شود.

۱.۵.۳ شواهد تجربی از تاثیر پارامترهای دوربین در معیارهای متریک

تحلیل ارائه‌شده در ادامه نشان می‌دهد که تغییر سمت دوربین در یک جفت استریو، با وجود شباهت محتوایی دو تصویر، می‌تواند در معیارهای وابسته به مقیاس اثر متفاوتی ایجاد کند. برای سنجش این موضوع، شاخص خطای مطلق نسبی[3] و ریشهٔ میانگین مربعات[4] برای خروجی مدل متریک Depth Anything V2 (انکودر ViT-S و نقطهٔ وارسی Hypersim) روی هر دو نمای مجموعهٔ Middlebury محاسبه شد. واحد استنباط آماری، محتوای صحنه بود: نتایج دو نسخهٔ تصحیح‌شدهٔ perfect و imperfect هر صحنه ابتدا میانگین‌گیری شد و

[1] Monocular Depth Estimation
[2] Camera Intrinsic Parameters
[3] Absolute Relative Error (AbsRel)
[4] Root Mean Squared Error (RMSE)

سپس آزمون زوجی دوسویه روی ۲۳ صحنهٔ مستقل انجام گرفت.

بازبینی فایل‌های کالیبراسیون یک نکتهٔ مهم را نیز روشن کرد: در هر جفت استریو، مقادیر f_x و f_y دو دوربین دقیقاً برابرند و c_y نیز تغییری ندارد؛ تفاوت نظام‌مند ماتریس‌ها در c_x است (اختلاف دوربین ۱ منهای دوربین ۰ در بازهٔ ۷۷٫۲۱۵ تا ۸۰۹٫۱۹۵ پیکسل). بنابراین این آزمایش را نباید آزمون علّی تغییر f_y دانست، بلکه مقایسهٔ زوجی دو سمت دوربین با دیدگاه‌ها و نقاط اصلی متفاوت است.

مطابق جدول ۱.۳، میانگین اختلاف RMSE برابر ۰٫۰۳۷۰ متر بود و آزمون t زوجی این اختلاف را در آستانهٔ ۰٫۰۵ معنادار نشان داد ($t(۲۲) = ۲٫۰۹۳$، $p = ۰٫۰۴۸$، بازهٔ اطمینان بوت‌استرپ ۹۵ درصد: $[۰٫۰۰۵۴, ۰٫۰۷۲۷]$). در مقابل، اختلاف AbsRel معنادار نبود ($t(۲۲) = ۱٫۴۹۳$، $p = ۰٫۱۵۰$، بازهٔ اطمینان: $[-۰٫۰۰۳۲, ۰٫۰۲۳۲]$). آزمون ناپارامتری ویلکاکسون نیز نتیجهٔ RMSE را تأیید کرد ($p = ۰٫۰۱۹۶$)، در حالی که نتیجهٔ AbsRel مرزی بود ($p = ۰٫۰۵۲۲$). شکل ۲.۳ مقادیر زوجی و جهت ناهمگن اختلاف‌ها را نشان می‌دهد.

جدول ۱.۳: ممیزی بازتولیدپذیر تفاوت دو سمت دوربین در Middlebury. اختلاف به‌صورت دوربین ۰ منهای دوربین ۱ و بازه‌ها به‌صورت بوت‌استرپ زوجی ۹۵ درصد گزارش شده‌اند ($n = ۲۳$).

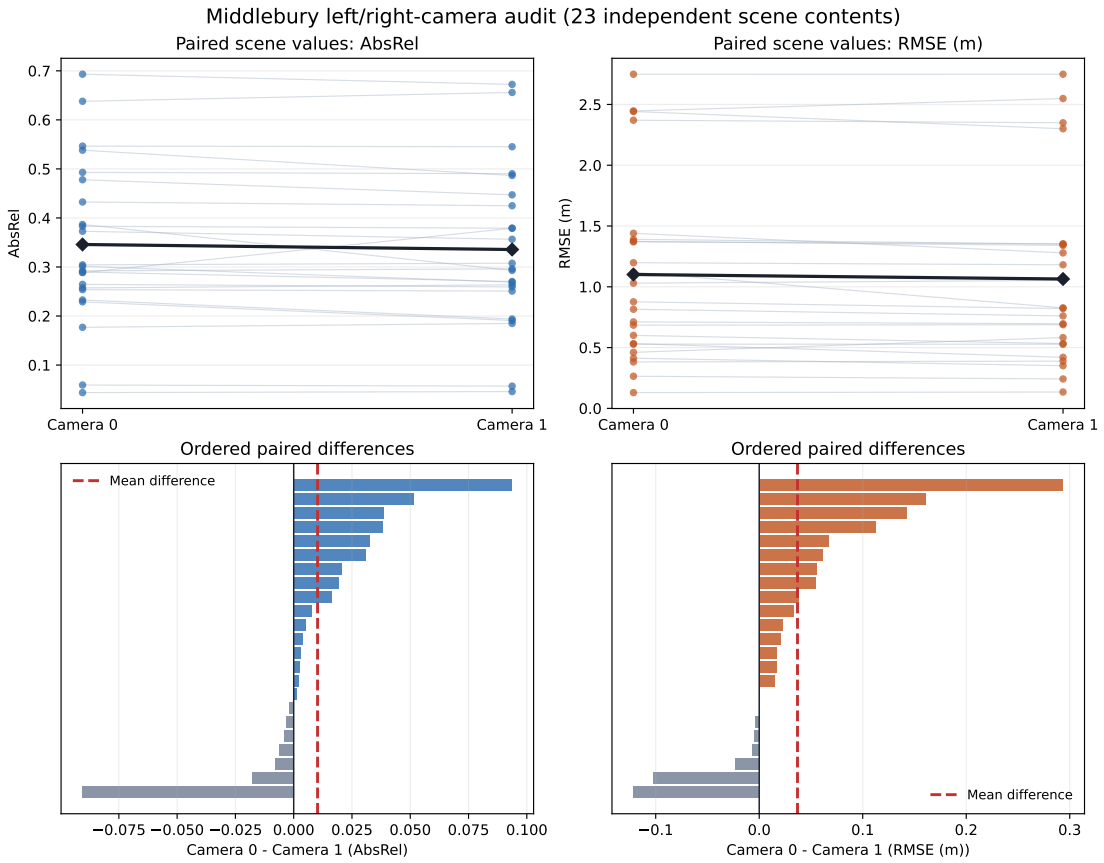
معیار	دوربین ۰ (میانگین ± انحراف معیار)	دوربین ۱ (میانگین ± انحراف معیار)	اختلاف میانگین [بازهٔ اطمینان]	$t(۲۲)$	p
↓AbsRel	۰٫۳۴۵۹ ± ۰٫۱۶۴۸	۰٫۳۳۵۶ ± ۰٫۱۶۳۴	۰٫۰۱۰۲ [−۰٫۰۰۳۲, ۰٫۰۲۳۲]	۱٫۴۹۳	۰٫۱۵۰
↓RMSE (متر)	۱٫۱۰۱۲ ± ۰٫۷۵۹۸	۱٫۰۶۴۲ ± ۰٫۷۶۰۷	۰٫۰۳۷۰ [۰٫۰۰۵۴, ۰٫۰۷۲۷]	۲٫۰۹۳	۰٫۰۴۸

این نتیجه از حساسیت بیشتر RMSE به تفاوت نظام‌مند میان دو سمت تصویربرداری پشتیبانی می‌کند، اما به‌تنهایی اثر علّی یک مؤلفهٔ مشخص از K را اثبات نمی‌کند؛ زیرا دو ورودی، دیدگاه و محتوای پیکسلی کاملاً یکسانی ندارند و مدل پایهٔ Depth Anything V2 نیز K را به‌صورت صریح دریافت نمی‌کند. شواهد علّی مصرف K در فصل بعد با کنترل تصویر و تغییر مستقیم ورودی K ارائه می‌شود.

۲.۵.۳ عدم کفایت رویکرد ضمنی نسبت به مشخصه‌های دوربین

روش‌هایی که بدون لحاظ صریح ماتریس مشخصهٔ دوربین[1] توسعه یافته‌اند، اغلب متکی بر راهبردهای یادگیری روابط نسبی یا بازسازی بر مبنای هندسهٔ ضمنی هستند؛ به این معنا که مدل، صرفاً براساس روابط داخلی پیکسل‌های تصویر، بدون آگاهی از مقیاس مطلق صحنه یا فواصل کانونی واقعی، اقدام به بازسازی عمق می‌کند. این سیاست، علاوه بر ایجاد وابستگی پنهان مدل نسبت به توزیع دوربین‌های موجود در مرحلهٔ آموزش، در عمل

[1] Camera Intrinsic Matrix (K)



شکل ۲.۳: مقایسهٔ زوجی دو سمت دوربین در Middlebury پس از میانگین‌گیری دو نسخهٔ تصحیح هر یک از ۲۳ محتوای صحنه. ردیف بالا مقادیر هر صحنه و میانگین کل (لوزی سیاه) و ردیف پایین اختلاف مرتب‌شدهٔ دوربین ۰ منهای دوربین ۱ را نشان می‌دهد. خط‌چین قرمز میانگین اختلاف است.

منجر به آن می‌شود که مدل به محض مواجهه با تصاویر ثبت‌شده توسط دوربینی با پارامترهای متفاوت، دیگر قادر به تولید تخمین‌های متریک دقیق نیست.

این موضوع به‌ویژه در کاربردهای عملی، که تصاویر ورودی ممکن است مجموعه‌ای ناهمگن از لحاظ نوع، رزولوشن، و پارامترهای اپتیکی باشند، مصداق می‌یابد و موجب کاهش جدی قابلیت اعتماد مدل در محیط‌های واقعی می‌شود. افزون بر آن، تضعیف ارزیابی مدل با معیارهای متریک (همانند RMSE) نه تنها مانعی برای ارتقاء کیفی سیستم‌های بینایی ماشینی محسوب می‌شود، بلکه حتی قابلیت رقابت مدل با سنسورهای عمق‌سنج فیزیکی را نیز با چالش جدی مواجه می‌سازد.

۳.۵.۳ کاربردهای عملی و انتقال‌پذیری به داده‌های جهان واقعی

در بسیاری از کاربردهای کلیدی تخمین عمق تک‌دوربینه، از جمله رباتیک، هدایت خودکار وسایل نقلیه، نقشه‌برداری سه‌بعدی، و واقعیت افزوده، داده‌های تصویری از منابع و دوربین‌های متعددی تأمین می‌شوند که هریک دارای ویژگی‌های اپتیکی منحصربه‌فردی هستند. انتقال موفق مدل از شرایط آزمایشگاهی به محیط‌های عملی، مستلزم آن است که سامانه بتواند برای هر تصویر ورودی، با استفاده از ماتریس مشخصهٔ دوربین متناظر آن، عمل تخمین عمق را به‌صورت سازگار و قابل اطمینانی انجام دهد. پژوهش‌هایی که اخیراً حول موضوع جبران و بهبود رفتار مدل نسبت به زاویهٔ دید (FOV) و کالیبراسیون دوربین در مجموعه‌داده‌های مختلف منتشر شده است [۱۲]، بر همین نکته تأکید دارند: لحاظ صریح مشخصات دوربین، کلید دستیابی به یک تخمین عمق متریک باکیفیت و مقاوم در کاربردهای واقعی است.

۴.۵.۳ لزوم طراحی مدل با ورود صریح پارامترهای درونی

برآیند این شواهد نظری و تجربی، این ضرورت را آشکار می‌کند که طراحی مدل‌هایی که به دنبال تخمین عمق متریک هستند، باید به‌طور ساختاری از اطلاعات پارامترهای دوربین (و به‌طور ویژه ماتریس K) بهره گیرند. لحاظ K، چه به‌صورت ورودی صریح به شبکه، چه از طریق تطبیق ساختار لایه‌های مدل با قیود فیزیکی تصویربرداری، می‌تواند پایداری و دقت تخمین عمق را در مواجهه با تنوع عملی داده‌ها تضمین کند.

رویکرد پیشنهادی ما در این پژوهش، مبتنی بر ورود ساخت‌یافتهٔ اطلاعات ماتریس مشخصهٔ دوربین در جریان یادگیری و استنتاج مدل است؛ رویکردی که شواهد ریاضی و تجربی، برتری آن را به‌ویژه در ارزیابی عمق متریک ثابت کرده‌اند. جزییات پیاده‌سازی این ایده و ادغام K با معماری پایه، در بخش‌های بعدی ارائه خواهد شد.

۶.۳ طراحی روش پیشنهادی: ادغام صریح ماتریس مشخصهٔ دوربین

بر اساس مباحث نظری ارائه‌شده در بخش‌های پیشین و شواهد تجربی حاصل از تحلیل حساسیت معیارهای متریک نسبت به پارامترهای درونی دوربین، در این پژوهش روشی برای ادغام صریح ماتریس مشخصهٔ دوربین در معماری Depth Anything V2 پیشنهاد می‌شود. ایدهٔ اصلی این روش، فراهم‌کردن امکان دسترسی مستقیم مدل به اطلاعات هندسی فرآیند تصویربرداری است، به‌گونه‌ای که شبکه به‌جای استنتاج ضمنی و ناپایدار این اطلاعات از روی داده، بتواند از قیود فیزیکی دوربین به‌صورت ساخت‌یافته بهره‌مند شود[۷، ۱۱].

در این بخش، ابتدا اصول کلی طراحی روش پیشنهادی تشریح می‌شود، سپس نحوهٔ نمایش و نرمال‌سازی ماتریس مشخصهٔ دوربین توضیح داده شده و در ادامه، راهبرد ادغام این اطلاعات با معماری encoder-decoder مدل پایهٔ Depth Anything V2 مورد بررسی قرار می‌گیرد.

۱.۶.۳ اصل طراحی: تفکیک اطلاعات تصویری و هندسی

در اغلب مدل‌های تخمین عمق تک‌دوربینه بدون نظارت، تصویر ورودی تنها منبع اطلاعاتی شبکه محسوب می‌شود و تمام قیود هندسی، اعم از فاصلهٔ کانونی یا میدان دید، به‌صورت ضمنی در وزن‌های مدل جذب می‌شوند[۱۴، ۶]. این رویکرد باعث می‌شود که مدل به توزیع دوربین‌های موجود در دادهٔ آموزشی وابسته شود و در مواجهه با داده‌هایی که از نظر پارامترهای درونی تفاوت دارند، عملکرد ناپایداری از خود نشان دهد.

روش پیشنهادی این پژوهش بر اصل تفکیک میان اطلاعات تصویری و اطلاعات هندسی استوار است. در این چارچوب، تصویر ورودی همچنان به‌عنوان منبع اطلاعات ظاهری و معنایی صحنه مورد استفاده قرار می‌گیرد، در حالی که ماتریس مشخصهٔ دوربین به‌عنوان توصیف‌گر هندسهٔ تصویربرداری، به‌صورت یک ورودی مستقل و مکمل در اختیار مدل قرار داده می‌شود. این تفکیک، به مدل اجازه می‌دهد تا نگاشت میان پیکسل و عمق را نه‌تنها بر اساس الگوهای بصری، بلکه با در نظر گرفتن مقیاس و هندسهٔ واقعی دوربین انجام دهد [۱۸، ۱۹].

۲.۶.۳ نمایش و پیش‌پردازش ماتریس مشخصهٔ دوربین

ماتریس مشخصهٔ دوربین به‌صورت زیر تعریف می‌شود:

$$K = \begin{bmatrix} f_x & ۰ & c_x \\ ۰ & f_y & c_y \\ ۰ & ۰ & ۱ \end{bmatrix}. \tag{۵.۳}$$

از آنجا که مقادیر این ماتریس به رزولوشن تصویر و تنظیمات فیزیکی دوربین وابسته‌اند، استفادهٔ مستقیم از مقادیر خام می‌تواند موجب ناپایداری عددی در فرآیند یادگیری شود [۲۰، ۷]. به همین دلیل، در روش پیشنهادی، پارامترهای درونی دوربین پیش از ورود به شبکه نرمال‌سازی می‌شوند. به‌طور خاص، مؤلفه‌های فاصلهٔ کانونی و

مختصات نقطهٔ اصلی بر حسب ابعاد تصویری که ماتریس K در آن بیان شده است نرمال می‌شوند:

$$\tilde{f}_x = \frac{f_x}{W}, \quad \tilde{f}_y = \frac{f_y}{H}, \quad \tilde{c}_x = \frac{c_x}{W}, \quad \tilde{c}_y = \frac{c_y}{H}, \tag{۶.۳}$$

که در آن W و H به‌ترتیب عرض و ارتفاع چارچوب مرجع پارامترهای دوربین هستند. این بازنمایی نرمال‌شده، امکان استفادهٔ پایدار از اطلاعات دوربین در مجموعه‌داده‌هایی با رزولوشن‌های متفاوت را فراهم می‌کند [۱۶، ۲۱].

افزون بر مؤلفه‌های نرمال‌شدهٔ فوق، برای آنکه بازنمایی ورودی به‌صورت مستقیم دربردارندهٔ مفهوم فیزیکی میدان دید[1] باشد، زوایای میدان دید افقی و عمودی نیز به‌صورت

$$\mathrm{fov}_w = ۲ \arctan\left(\frac{W}{۲ f_x}\right), \qquad \mathrm{fov}_h = ۲ \arctan\left(\frac{H}{۲ f_y}\right) \tag{۷.۳}$$

محاسبه می‌شوند. مطابق پیاده‌سازی نهایی، زاویه‌ها بر π، میانگین فاصلهٔ کانونی بر بزرگ‌ترین بُعد تصویر و مساحت تصویر بر مساحت مرجع ۱۰۸۰ × ۱۹۲۰ تقسیم می‌شوند. بنابراین، بردار توصیف‌گر هندسی دوربین دقیقاً به‌صورت

$$\mathbf{k} = \left[\frac{f_x}{W}, \frac{f_y}{H}, \frac{c_x}{W}, \frac{c_y}{H}, \frac{\mathrm{fov}_h}{\pi}, \frac{\mathrm{fov}_w}{\pi}, \frac{W}{H}, \frac{f_x + f_y}{۲ \max(H, W)}, \frac{HW}{۱۹۲۰ \times ۱۰۸۰}\right] \tag{۸.۳}$$

تعریف می‌شود و به‌عنوان ورودی هندسی مدل مورد استفاده قرار می‌گیرد. جدول ۲.۳ معنای هندسی هر یک از نه مؤلفهٔ این بردار را تشریح می‌کند.

نکتهٔ اساسی در این بازنمایی آن است که ماتریس K و ابعاد (H, W) باید هر دو در یک چارچوب مختصات یکسان بیان شوند؛ در غیر این صورت، کدگذاری حاصل از منظر هندسی نادرست خواهد بود. اهمیت عملی این نکته در بخش جزئیات پیاده‌سازی و در تحلیل اشکال‌زدایی فصل چهارم آشکار می‌شود.

۳.۶.۳ راهبرد ادغام اطلاعات دوربین با معماری Depth Anything V2

برای ادغام بردار پارامترهای درونی دوربین با معماری پایه، در این پژوهش از یک واحد کدگذار دوربین[2] استفاده می‌شود که وظیفهٔ آن، تبدیل بردار کم‌بعد $\mathbf{k}$ به یک توکن دوربین سازگار با فضای بازنمایی ترنسفورمر است. این واحد شامل یک شبکهٔ تمام‌متصل دولایه برای نگاشت اولیه و سپس چند بلوک ترنسفورمر برای پالایش

[1] Field of View (FoV)
[2] Camera Encoder

جدول ۲.۳: مؤلفه‌های بردار توصیف‌گر هندسی دوربین $\mathbf{k} \in \mathbb{R}^{۹}$ و نقش هندسی هر یک. تمامی مؤلفه‌ها بی‌بعد و در مرتبهٔ واحد هستند تا فرآیند یادگیری از نظر عددی پایدار بماند.

شمارهٔ مؤلفه	تعریف	نقش هندسی
۱	f_x/W	فاصلهٔ کانونی افقی، مستقل از رزولوشن
۲	f_y/H	فاصلهٔ کانونی عمودی، مستقل از رزولوشن
۳	c_x/W	موقعیت افقی نقطهٔ اصلی؛ به جابه‌جایی مرکز تصویر در برش حساس است
۴	c_y/H	موقعیت عمودی نقطهٔ اصلی
۵	fov_h/π	میدان دید عمودی به‌صورت زاویه‌ای و نرمال‌شده
۶	fov_w/π	میدان دید افقی به‌صورت زاویه‌ای و نرمال‌شده
۷	W/H	نسبت ابعاد قاب تصویر
۸	$\dfrac{f_x + f_y}{۲\max(H, W)}$	کانون میانگین نسبت به بزرگ‌ترین بُعد؛ سنجه‌ای فشرده از «تنگی» میدان دید
۹	$\dfrac{HW}{۱۹۲۰ \times ۱۰۸۰}$	مساحت قاب نسبت به یک چارچوب مرجع ثابت

بازنمایی است که خروجی آن یک توکن با ابعاد برابر با بعد نهان انکودر بینایی است:

$$\mathbf{z}_k = \phi(\mathbf{k}), \tag{۹.۳}$$

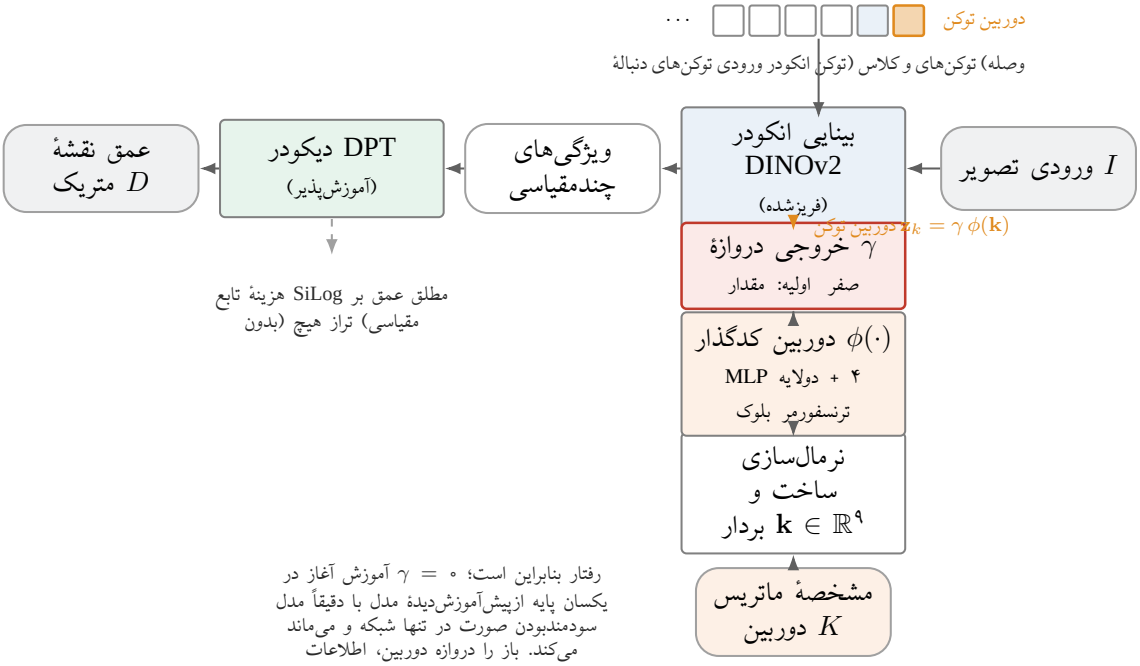
که در آن $\phi(\cdot)$ تابع نگاشت پارامتری قابل یادگیری است.

توکن دوربین $\mathbf{z}_k$ سپس در کنار توکن‌های تصویری، به دنبالهٔ توکن‌های انکودر مبتنی بر DINOv2 تزریق می‌شود؛ به‌گونه‌ای که سازوکار خودتوجهی بتواند در تمامی لایه‌های پس از نقطهٔ تزریق، اطلاعات هندسی دوربین را با بازنمایی‌های بصری ترکیب کند. لایهٔ تزریق از طریق پارامتر قابل تنظیم `--cam-token-inject-layer` کنترل می‌شود و در تنظیمات پیش‌فرض، تزریق از نخستین لایهٔ انکودر انجام می‌گیرد. این راهبرد با روش‌های شرطی‌سازی مبتنی بر توکن در معماری‌های ترنسفورمری و نیز رویکردهای تنظیم ویژگی به‌کمک بردارهای شرطی نظیر FiLM هم‌خانواده است [۲۲]؛ با این تفاوت که در اینجا، شرطی‌سازی از طریق سازوکار توجه انجام می‌شود و نیازی به دستکاری مستقیم نگاشت‌های ویژگی نیست.

نکتهٔ طراحی مهم در این ادغام، استفاده از یک دروازهٔ خروجی با مقداردهی اولیهٔ صفر[1] است: خروجی کدگذار دوربین در ابتدای آموزش در یک ضریب قابل یادگیری با مقدار اولیهٔ صفر ضرب می‌شود. در نتیجه، در گام‌های نخست آموزش، توکن دوربین عملاً برداری صفر است و رفتار مدل دقیقاً با مدل از پیش‌آموزش‌دیدهٔ پایه

[1] Zero-initialized Output Gate

یکسان باقی می‌ماند؛ سپس در طول آموزش، شبکه به‌تدریج و تنها در صورتی که اطلاعات دوربین برای کاهش خطا مفید باشد، این دروازه را باز می‌کند. این سازوکار، از آسیب رساندن به بازنمایی‌های ازپیش‌آموخته در آغاز فاین‌تیون جلوگیری کرده و پایداری فرآیند آموزش را تضمین می‌کند [۱۸، ۲۳].

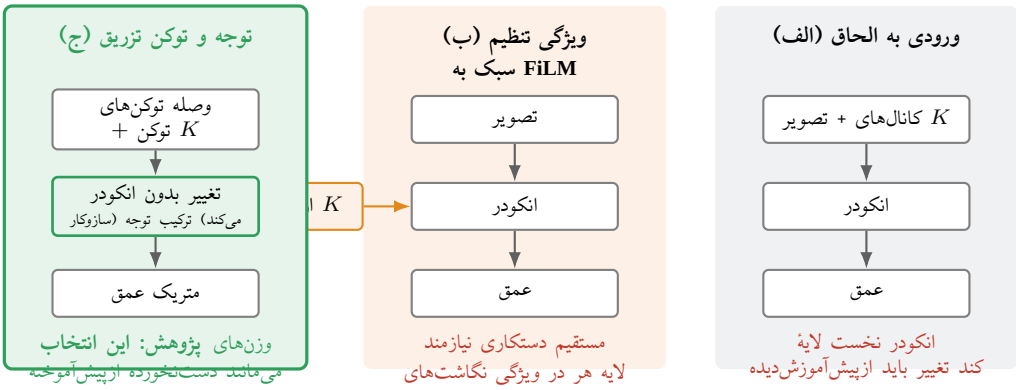


شکل ۳.۳: معماری مدل پیشنهادی (دارای K). مسیر بالایی، مسیر بصری استاندارد مدل پایهٔ Depth Anything V2 است: تصویر ورودی توسط انکودر ازپیش‌آموزش‌دیدهٔ DINOv2 (فریزشده) به ویژگی‌های چندمقیاسی تبدیل و سپس توسط دیکودر DPT به نقشهٔ عمق بازگردانده می‌شود. مسیر پایینی، افزودهٔ این پژوهش است: ماتریس K پس از نرمال‌سازی به بردار نه‌بعدی $\mathbf{k}$ (جدول ۲.۳) تبدیل، توسط کدگذار دوربین به یک توکن نهان نگاشت، و پس از عبور از یک دروازهٔ خروجی با مقداردهی اولیهٔ صفر، به ابتدای دنبالهٔ توکن‌های انکودر تزریق می‌شود. از آن لایه به بعد، سازوکار خودتوجهی می‌تواند اطلاعات هندسی را با بازنمایی‌های بصری ترکیب کند. مدل کنترل (فاقد K) دقیقاً همین معماری بدون شاخهٔ پایینی است.

شکل ۴.۳ این انتخاب طراحی را در برابر دو گزینهٔ متداول دیگر قرار می‌دهد.

۴.۶.۳ مزایای مورد انتظار از ادغام صریح پارامترهای درونی

ادغام صریح ماتریس مشخصهٔ دوربین، چند مزیت کلیدی را به همراه دارد. نخست آنکه مدل قادر خواهد بود نگاشت میان پیکسل و عمق را با در نظر گرفتن مقیاس واقعی دوربین انجام دهد و از این طریق، دقت تخمین عمق متریک افزایش یابد. دوم آنکه وابستگی مدل به توزیع خاصی از دوربین‌ها در دادهٔ آموزشی کاهش می‌یابد و تعمیم‌پذیری آن در مواجهه با داده‌های چنددوربینه بهبود پیدا می‌کند. این موضوع به‌ویژه با توجه به نتایج تجربی



شکل ۴.۳: مقایسهٔ سه سازوکار ممکن برای شرطی‌سازی مدل بر پارامترهای درونی دوربین. (الف) الحاق K به‌صورت کانال‌های اضافی به تصویر ورودی، لایهٔ نخست انکودر ازپیش‌آموزش‌دیده را تغییر می‌دهد و بخشی از دانش اولیه را از بین می‌برد. (ب) تنظیم ویژگی به سبک FiLM [۲۲] نیازمند دخالت مستقیم در نگاشت‌های ویژگی است. (ج) راهبرد این پژوهش: افزودن یک توکن به دنبالهٔ ورودی، به‌گونه‌ای که هستهٔ انکودر دست‌نخورده بماند و ترکیب اطلاعات هندسی و بصری بر عهدهٔ خودِ سازوکار توجه باشد. این انتخاب، در کنار دروازهٔ خروجی صفرمقدار اولیه، آموزش با انکودر کاملاً فریزشده را ممکن می‌سازد.

ارائه‌شده در تحلیل RMSE در بخش پیشین، اهمیت ویژه‌ای دارد. در نهایت، این طراحی امکان توسعهٔ مدل به سناریوهای پیشرفته‌تر، نظیر یادگیری انتقالی میان مجموعه‌داده‌های با میدان دید متفاوت، را نیز فراهم می‌کند.

در بخش‌های بعدی، نحوهٔ آموزش این مدل، تابع هزینهٔ مورد استفاده، و تنظیمات تجربی برای ارزیابی اثر ادغام پارامترهای درونی دوربین به‌تفصیل تشریح خواهد شد.

۷.۳ راهبرد آموزش مدل پیشنهادی (دارای K)

در این پژوهش، آموزش مدل پیشنهادی (دارای K) به‌صورت فاین‌تیون[1] یک مدل ازپیش‌آموزش‌یافته انجام شده است. به‌طور خاص، از مدل پایهٔ Depth Anything V2، نسخهٔ غیرمتریک با وزن‌های اولیهٔ مدل Large به‌عنوان نقطهٔ شروع استفاده شده و فرآیند آموزش با هدف تبدیل آن به یک مدل متریک پایدار، با لحاظ صریح پارامترهای درونی دوربین، طراحی شده است. این راهبرد، ضمن بهره‌گیری از دانش غنی مدل معلم، امکان انطباق مدل با تنظیمات هندسی جدید را بدون نیاز به آموزش از ابتدا فراهم می‌کند.

برای رفع ابهام اصطلاح‌شناختی، عنوان کلی پژوهش به تخمین عمق بدون نظارت و استفاده از یک مدل بنیادین ازپیش‌آموخته اشاره دارد، اما مرحلهٔ نهایی فاین‌تیون کنترل‌شده در این پایان‌نامه نظارتی است و از نگاشت‌های عمق مرجع NYUv2 استفاده می‌کند. تقطیر دانش در زیرساخت پیاده‌سازی شده، ولی در

[1] Fine-tuning

آزمایش‌های نهایی غیرفعال است. این تفکیک ضروری است تا بهبود مشاهده‌شده به ورودی K و برنامهٔ فاین‌تیون نسبت داده شود، نه به یک سیگنال آموزشی جانبی.

۱.۷.۳ مجموعه‌داده‌های مورد استفاده

در مرحلهٔ اکتشافی، ترکیب‌هایی از Virtual KITTI[1]، NYUv2[2]، DIODE[3]، Cityscapes و Middlebury آزموده شد. این داده‌ها از نظر نوع صحنه، شرایط تصویربرداری و مشخصات دوربین متنوع‌اند [۲۴، ۲۵، ۲۶]؛ بااین‌حال، نتایج فصل چهارم نشان داد که فاین‌تیون چندمجموعه‌ای محدود، تعمیم صفرنمونهٔ مدل بنیادین را تخریب می‌کند. ازاین‌رو، مطالعهٔ نهایی برای جداسازی علّی سهم K به NYUv2 محدود شد و هر زوج مدل پیشنهادی (دارای K)/مدل کنترل (فاقد K) دقیقاً داده‌ها و برنامهٔ آموزشی یکسانی دریافت کرد.

جدول ۳.۳ مشخصات این مجموعه‌داده‌ها و نقش هر یک در مراحل مختلف پژوهش را خلاصه می‌کند.

جدول ۳.۳: مجموعه‌داده‌های به‌کاررفته، ویژگی‌های تصویربرداری آن‌ها و نقش هر یک در مراحل اکتشافی و نهایی پژوهش. ستون «دسترسی به K» تعیین‌کنندهٔ آن است که مجموعه‌داده برای آزمایش مدل پیشنهادی (دارای K) قابل استفاده باشد.

مجموعه‌داده	نوع محیط	منبع عمق مرجع	دسترسی به K	نقش در این پژوهش
NYUv2 [۲۵]	داخلی، واقعی	حسگر Kinect	دارد، ثابت برای کل مجموعه	**بستر مطالعهٔ کنترل‌شده‌نهایی**؛ تقسیم بدون نشت و پروتکل استاندارد Eigen
Virtual KITTI [۲۴]	بیرونی، مصنوعی	رندر دقیق موتور گرافیکی	دارد، دقیق	مرحلهٔ اکتشافی؛ سنجش پایداری بین‌دیتاستی
KITTI	بیرونی، واقعی	لایدار، پراکنده	دارد، از فایل کالیبراسیون	محک ارزیابی صفرنمونه و راستی‌آزمایی زنجیرهٔ ارزیابی
DIODE [۲۶]	داخلی و بیرونی	اسکنر لیزری	دارد	مرحلهٔ اکتشافی؛ تنوع بالای صحنه
Cityscapes	شهری، واقعی	تطبیق استریو	دارد	مطالعهٔ موردی رفتار خطا در دامنهٔ عمق بزرگ
Middlebury	داخلی، واقعی	استریوی دقت‌بالا	دارد، دو نمای کالیبره‌شده	ممیزی اثر تفاوت دو سمت دوربین (جدول ۱.۳)
Hypersim	داخلی، مصنوعی	رندر دقیق	دارد	نقطهٔ وارسی متریک مرجع مدل پایهٔ Depth Anything V2

داده‌های مورد استفاده عمدتاً به‌صورت تک‌تصویر یا توالی‌های کوتاه از تصاویر ثبت‌شده‌اند، با این تفاوت که در فرآیند آموزش حاضر، هر تصویر به‌صورت مستقل پردازش شده و هیچ‌گونه قید زمانی یا وابستگی صریح میان

[1] Virtual KITTI
[2] NYUv2
[3] DIODE

فریم‌های متوالی اعمال نشده است. این انتخاب، با هدف تمرکز بر مسئلهٔ تخمین عمق تک‌تصویر و حذف اثرات جانبی ناشی از حرکت دوربین یا برآورد ژست صورت گرفته است.

در مواردی که اطلاعات ماتریس مشخصهٔ دوربین به‌صورت صریح در داده‌ها موجود بوده، این اطلاعات مستقیماً مورد استفاده قرار گرفته و در سایر موارد، پارامترهای دوربین مطابق با تنظیمات مرجع هر مجموعه‌داده استخراج یا بازسازی شده‌اند.

۲.۷.۳ تفکیک صحیح مجموعه‌های آموزش و آزمون در NYUv۲

نکتهٔ روش‌شناختی مهمی که در جریان این پژوهش شناسایی و اصلاح شد، نحوهٔ تفکیک داده‌های آموزش و آزمون در مجموعه‌دادهٔ NYUv2 است. پروتکل استاندارد ارزیابی در ادبیات این حوزه، زیرمجموعهٔ آزمون ۶۵۴ تصویری موسوم به تقسیم‌بندی Eigen است [۲] که مدل Depth Anything V2 نیز نتایج خود را بر همین مبنا گزارش می‌کند [۱۳]. در نسخهٔ اولیهٔ زیرساخت آموزش این پژوهش، تقسیم‌بندی آموزش/اعتبارسنجی به‌صورت تصادفی بر روی کل ۱۴۴۹ تصویر مجموعهٔ برچسب‌دار انجام می‌شد؛ بررسی دقیق نشان داد که با این رویه، ۵۲۷ تصویر از ۶۵۴ تصویر آزمون استاندارد (معادل ۸۰٫۶ درصد) در مجموعهٔ آموزش قرار می‌گیرند و در نتیجه هرگونه ارزیابی روی مجموعهٔ آزمون، دچار نشت داده[1] و به‌صورت مصنوعی خوش‌بینانه خواهد بود.

برای رفع این مشکل، فرآیند تقسیم‌بندی به‌گونه‌ای اصلاح شد که ابتدا تمامی ۶۵۴ تصویر آزمون استاندارد به‌طور کامل کنار گذاشته شوند و سپس تقسیم ۲۰/۸۰ آموزش/اعتبارسنجی تنها بر روی ۷۹۵ تصویر باقی‌مانده انجام گیرد (۶۳۶ تصویر آموزش و ۱۵۹ تصویر اعتبارسنجی). عدم هم‌پوشانی سه مجموعه نیز با آزمون خودکار راستی‌آزمایی شده است. تمامی نتایج نهایی گزارش‌شده در فصل چهارم بر مبنای این تفکیک بدون نشت به‌دست آمده‌اند.

۳.۷.۳ مدل پایهٔ Depth Anything V2 و راهبرد فاین‌تیون

نسخهٔ غیرمتریک مدل پایهٔ Depth Anything V2 به‌صورت distillation teacher-student[2] آموزش دیده و دارای توانایی بالایی در استخراج ویژگی‌های معنایی و هندسی تصویر است. در این پژوهش، این مدل به‌عنوان دانش اولیه در نظر گرفته شده و فرآیند فاین‌تیون با هدف افزودن قابلیت تخمین عمق متریک و تطبیق با اطلاعات دوربین انجام شده است.

[1] Data Leakage
[2] Teacher–Student Distillation

برای حفظ پایداری آموزش و جلوگیری از بیش‌برازش[1]، بخش encoder مبتنی بر DINOv2[2] در طول آموزش فریز شده است. این تصمیم مبتنی بر این فرض است که ویژگی‌های استخراج‌شده توسط DINOv2، پیشاپیش از غنای کافی برای نمایش محتوای تصویری برخوردارند و نیازی به بازآموزی در این مرحله ندارند [۲۷]. در مقابل، بخش decoder و همچنین ماژول‌های مرتبط با تخمین عمق و پردازش پارامترهای دوربین، به‌صورت کامل قابل یادگیری باقی مانده‌اند.

۴.۷.۳ فعال‌سازی پشتیبانی از پارامترهای دوربین

گزینهٔ `--use-camera-intrinsics` در اسکریپت آموزش، استفاده از ماتریس مشخصهٔ دوربین را در فرآیند آموزش فعال می‌کند. با فعال‌سازی این گزینه، بردار نرمال‌سازی‌شدهٔ پارامترهای دوربین مطابق با توضیحات بخش ۳-۶ استخراج شده و از طریق یک encoder اختصاصی به فضای نهان مدل نگاشت می‌شود.

علاوه بر این، امکان کنترل محل تزریق اطلاعات دوربین در معماری شبکه از طریق پارامتر `--cam-token-inject-layer` فراهم شده است. این پارامتر لایهٔ تزریق توکن به انکودر DINOv2 را تعیین می‌کند. در تمام آزمایش‌های گزارش‌شده مقدار پیش‌فرض، یعنی لایهٔ نخست انکودر، استفاده شده است؛ حساسیت نسبت به محل تزریق و تزریق چندلایه در این پژوهش پیمایش نشده و به‌عنوان مسیر کار آینده گزارش می‌شود.

۵.۷.۳ تنظیمات آموزش و بهینه‌سازی

فرآیند آموزش با اسکریپت `train.py` و در دو حالت انجام شده است: `--model-type basic` برای خروجی نسبی در فضای عمق معکوس، و `--model-type metric` برای عمق مطلق بر حسب متر. حالت نسبی برای بازتولید پروتکل استاندارد و تشخیص محدودیت تراز مقیاس به‌کار رفت؛ حالت متریک، بستر اصلی آزمون فرضیهٔ نقش K در بازیابی مقیاس بود.

به‌منظور سازگاری بهتر فرآیند فاین‌تیون با وزن‌های ازپیش‌آموزش‌یافته، نرخ یادگیری بخش‌های مختلف شبکه به‌صورت ناهمسان تنظیم شده است. به‌طور خاص، برای لایه‌های مربوط به تخمین عمق و encoder پارامترهای

[1] Overfitting
[2] DINOv2

دوربین، از ضریب تقویت نرخ یادگیری استفاده شده است:

$$\text{–head-lr-multiplier} \tag{۱۰.۳}$$

این راهبرد باعث می‌شود که لایه‌های جدید یا تغییر یافته، سریع‌تر با هدف متریک آموزش تطبیق یابند، در حالی که سایر بخش‌های شبکه دچار نوسان شدید نشوند.

همچنین، برای افزایش پایداری عددی آموزش، از برش گرادیان[1] و در برخی تنظیمات از warm-up تدریجی نرخ یادگیری استفاده شده است. این تنظیمات به‌ویژه در مراحل ابتدایی آموزش، از واگرایی گرادیان‌ها جلوگیری کرده و همگرایی نرم‌تری را فراهم می‌کنند.

۶.۷.۳ فضای هدف آموزش برای مدل‌های نسبی: عمق معکوس

یکی از یافته‌های روش‌شناختی مهم این پژوهش، اهمیت هم‌خوانی فضای هدف آموزش با فضای خروجی مدل از پیش‌آموزش‌دیده است. مدل پایهٔ Depth Anything V2، نسخهٔ نسبی، مطابق با سنت روش‌های خانوادهٔ MiDaS [۱۱]، در فضای عمق معکوس (دیسپاریتی)[2] آموزش دیده و خروجی آن با عمق رابطهٔ وارون دارد: نقاط نزدیک مقادیر بزرگ و نقاط دور مقادیر کوچک دریافت می‌کنند.

در نسخهٔ اولیهٔ زیرساخت آموزش این پژوهش، مقادیر عمق مستقیم به‌عنوان هدف آموزشی مدل نسبی به‌کار می‌رفت. تحلیل دقیق این تنظیم نشان داد که ترکیب آن با تابع هزینهٔ تغییرناپذیر نسبت به مقیاس و جابجایی، به یک شکست خاموش[3] منجر می‌شود: از یک سو، محدودسازی ضریب مقیاس s به مقادیر مثبت در پیاده‌سازی تابع هزینه، برازش رابطهٔ وارون میان خروجی دیسپاریتی و هدف عمق مستقیم را ناممکن می‌سازد؛ و از سوی دیگر، مؤلفهٔ تطابق گرادیان که بر روی مقادیر ترازنشده محاسبه می‌شود، دامنهٔ خروجی مدل را به‌شدت به سمت صفر سوق می‌دهد. برآیند این دو عامل آن است که بهینهٔ عملی تابع هزینه، خروجی ثابت صفر است و تابع فعال‌سازی ReLU در انتهای سر شبکه، این حالت را برگشت‌ناپذیر می‌کند (گرادیان صفر). آزمایش کنترل‌شدهٔ ابزارگذاری‌شده نشان داد که مدل ViT-L تنها در حدود ۱۸ گام آموزشی به‌طور کامل به این حالت فرو می‌ریزد؛ این در حالی است که منحنی هزینهٔ آموزش ظاهری کاملاً طبیعی دارد، زیرا با خروجی ثابت، مقدار هزینه صرفاً تابع آماره‌های داده در هر دسته است. جزئیات تجربی این پدیده و فرآیند تشخیص آن در فصل چهارم گزارش شده است.

[1] Gradient Clipping
[2] Inverse Depth / Disparity
[3] Silent Failure

بر این اساس، در چارچوب نهایی آموزش، هدف مدل‌های نسبی به فضای عمق معکوس منتقل شد:

$$T_{\text{disp}} = \frac{۱}{\max(D^{GT}, d_{\min})}, \tag{۱۱.۳}$$

که در آن $d_{\min}$ کران پایین عمق معتبر است. با این تغییر، فضای هدف با فضای خروجی مدل ازپیش‌آموزش‌دیده هم‌راستا شده و فرآیند فاین‌تیون بدون فروریزش و با پایداری کامل انجام می‌گیرد. متناظر با آن، ارزیابی اعتبارسنجی حین آموزش نیز با ترازکردن مقیاس و جابجایی در فضای عمق معکوس و سپس وارون‌سازی به فضای عمق انجام می‌شود و نوع فضای هدف در فراداده‌های ذخیره‌شدهٔ هر نقطهٔ وارسی[1] ثبت می‌گردد تا زنجیرهٔ ارزیابی به‌صورت خودکار رفتار صحیح را اتخاذ کند.

۷.۷.۳ افزونه‌سازی هندسی آگاه از پارامترهای دوربین

چالش بنیادی در آموزش مدل پیشنهادی (دارای K) بر روی یک مجموعه‌دادهٔ تک‌دوربینه (نظیر NYUv2 که تمامی تصاویر آن با یک حسگر Kinect ثبت شده‌اند) آن است که بردار $\mathbf{k}$ برای تمامی نمونه‌ها ثابت است و در نتیجه هیچ سیگنال تمایزبخشی برای واحد کدگذار دوربین فراهم نمی‌کند. افزون بر این، بررسی زیرساخت اولیهٔ آموزش نشان داد که برش تصادفی[2] متداول در خط لولهٔ پیش‌پردازش، مرکز تصویر مؤثر را جابه‌جا می‌کند بدون آنکه ماتریس K متناظراً به‌روزرسانی شود؛ در نتیجه، کدگذار دوربین در عمل اطلاعاتی ثابت و از منظر هندسی نادرست دریافت می‌کرد.

برای رفع هم‌زمان هر دو مشکل، در این پژوهش یک راهبرد افزونه‌سازی هندسی دقیق طراحی و پیاده‌سازی شد که در آن، برای هر نمونهٔ آموزشی، پنجره‌ای تصادفی با نسبت ابعادی ثابت ۳:۴ و ضریب بزرگ‌نمایی s (در بازهٔ ۱ تا حداکثر ۱٫۷ یا ۲٫۲ بسته به تنظیم آزمایش، و با احتمال ۲۰ درصد دقیقاً قاب کامل) از تصویر بریده می‌شود و ماتریس K به‌صورت دقیق در تمامی مراحل هندسی به‌روزرسانی می‌گردد: جابه‌جایی نقطهٔ اصلی متناسب با مبدأ پنجرهٔ برش،

$$c'_x = c_x - x_\circ, \qquad c'_y = c_y - y_\circ, \tag{۱۲.۳}$$

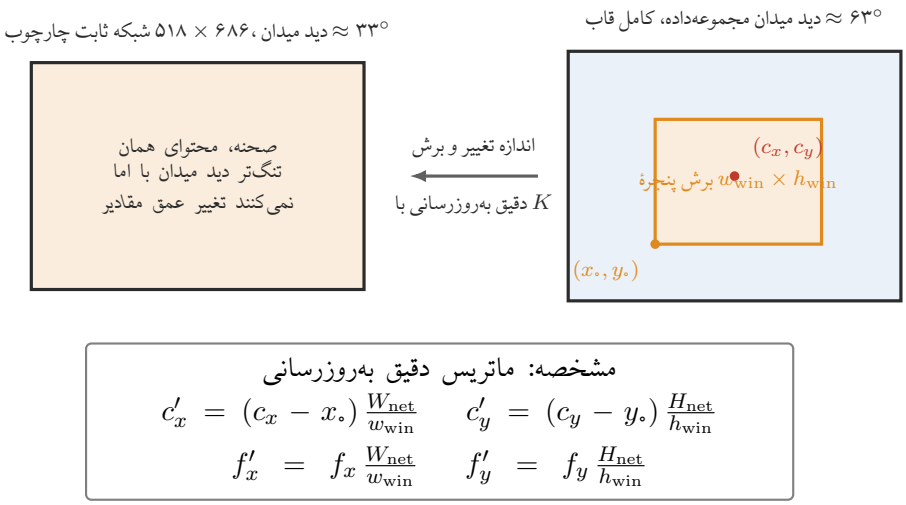
[1] Checkpoint
[2] Random Crop

و مقیاس‌دهی فاصلهٔ کانونی و نقطهٔ اصلی متناسب با تغییر اندازهٔ پنجره به چارچوب ثابت ورودی شبکه:

$$\begin{aligned} f'_x &= f_x \frac{W_{\text{net}}}{w_{\text{win}}}, \quad c'_x = (c_x - x_\circ)\frac{W_{\text{net}}}{w_{\text{win}}}, \\ f'_y &= f_y \frac{H_{\text{net}}}{h_{\text{win}}}, \quad c'_y = (c_y - y_\circ)\frac{H_{\text{net}}}{h_{\text{win}}}. \end{aligned} \tag{۱۳.۳}$$

که در آن $(H_{\text{net}}, W_{\text{net}})$ ابعاد ثابت چارچوب ورودی شبکه (در پیاده‌سازی حاضر ۶۸۶ × ۵۱۸، مضرب اندازهٔ وصله‌های ترنسفورمر) است. مقادیر عمق در این تبدیل تغییر نمی‌کنند، زیرا نقاط صحنه همان نقاط پیشین هستند.

شکل ۵.۳ این زنجیره را به‌صورت شماتیک نشان می‌دهد.



شکل ۵.۳: افزونه‌سازی هندسی آگاه از پارامترهای دوربین. یک پنجرهٔ تصادفی با نسبت ابعاد ثابت ۴:۳ از قاب کامل بریده و به چارچوب ثابت ورودی شبکه تغییر اندازه داده می‌شود؛ هم‌زمان، نقطهٔ اصلی و فاصلهٔ کانونی مطابق روابط پایین شکل به‌صورت دقیق به‌روزرسانی می‌گردند. نتیجه، شبیه‌سازی دوربین‌هایی با میدان دید متفاوت از یک صحنهٔ واحد است. در زیرساخت اولیهٔ پژوهش، همین برش انجام می‌شد بدون آنکه K اصلاح شود؛ در نتیجه کدگذار دوربین ورودی هندسی نادرست دریافت می‌کرد (بخش ۸.۹.۳).

وارونه‌سازی افقی نیز هم‌زمان بر تصویر، نگاشت عمق، ماسک اعتبار و ماتریس مشخصه اعمال می‌شود. در این حالت، مختصهٔ افقی نقطهٔ اصلی مطابق $c'_x = W - c_x$ اصلاح می‌گردد. همچنین ابعاد چارچوبی که K در آن تعریف شده است همراه هر نمونه نگهداری می‌شود تا نرمال‌سازی مؤلفه‌های دوربین در آموزش و استنتاج دقیقاً در یک دستگاه مختصات انجام شود. این کنترل‌ها ضروری‌اند، زیرا تغییر تصویر بدون تبدیل متناظر K، حتی در صورت اجرای بی‌خطای برنامه، ورودی هندسی ناسازگار تولید می‌کند.

از منظر هندسی، این افزونه‌سازی معادل شبیه‌سازی دوربین‌هایی با میدان دید متفاوت از یک صحنهٔ واحد است: بزرگ‌نمایی s میدان دید مؤثر را از حدود ۶۳ درجه (قاب کامل NYUv2) تا حدود ۳۳ درجه تغییر می‌دهد.

بدین ترتیب، مدل پیشنهادی (دارای K) در طول آموزش با زوج‌های (تصویر، K) مواجه می‌شود که در آن‌ها ابهام میان مقیاس صحنه و میدان دید به‌صورت واقعی وجود دارد و تنها از طریق اطلاعات K قابل رفع است؛ در حالی که مدل کنترل (فاقد K)، دقیقاً همین تصاویر افزونه‌سازی‌شده را بدون اطلاعات K دریافت می‌کند. این طراحی، مبنای مطالعهٔ کنترل‌شدهٔ فصل چهارم برای سنجش سهم واقعی پارامترهای درونی دوربین است.

۸.۷.۳ افزونه‌سازی لرزش کانونی: غیرقابل‌حل‌کردن مقیاس بدون K

افزونه‌سازی زوم‌ـبرش زیربخش قبل، میدان دید را متنوع می‌کند، اما یک محدودیت بنیادی دارد: در تصاویر معمولی، نشانه‌های معنایی صحنه (ابعاد آشنای اشیاء، مبلمان، درها) به‌تنهایی برای بازیابی مقیاس تقریبی کافی‌اند و شبکه می‌تواند بدون مراجعه به K نیز پاسخ قابل قبولی بدهد. برای آنکه اتکای مدل به پارامترهای درونی دوربین به یک ضرورت تبدیل شود، در این پژوهش افزونه‌سازی مکملی با عنوان «لرزش کانونی»[1] طراحی شد که مستقیماً از معادلهٔ تصویربرداری روزنه‌ای الهام می‌گیرد. طبق این معادله، $u = f \cdot X/Z$، اگر فاصلهٔ کانونی در ضریب α و هم‌زمان تمام عمق‌های صحنه نیز در همان α ضرب شوند، تصویر حاصل دقیقاً یکسان باقی می‌ماند. بر همین اساس، برای هر نمونهٔ آموزشی ضریبی به‌صورت لگاریتمی‌ـیکنواخت از بازهٔ

$$\alpha \sim \exp\big(\mathcal{U}(-\ln m, +\ln m)\big), \qquad m = ۱/۶, \tag{۱۴.۳}$$

نمونه‌برداری شده و تبدیل

$$f_x \mapsto \alpha f_x, \qquad f_y \mapsto \alpha f_y, \qquad Z \mapsto \alpha Z \tag{۱۵.۳}$$

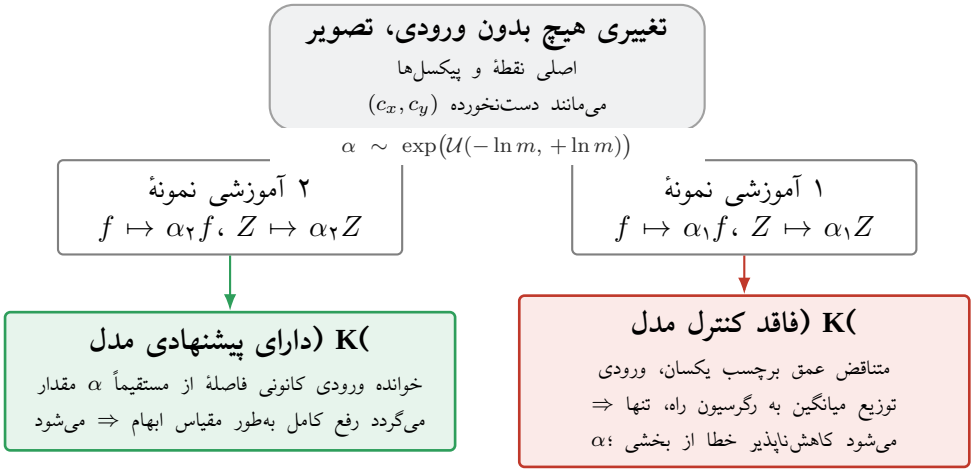
اعمال می‌شود؛ پیکسل‌های تصویر و مختصات نقطهٔ اصلی (c_x, c_y) دست‌نخورده می‌مانند. این تبدیل از منظر هندسی دقیق است (برخلاف بازنمونه‌برداری، هیچ خطای درون‌یابی ایجاد نمی‌کند) و هیچ نشتی اطلاعاتی ندارد، زیرا تصویر ورودی تغییری نمی‌کند.

پیامد این طراحی برای دو بازوی مطالعهٔ کنترل‌شده متقارن نیست: مدل کنترل (فاقد K) با تصاویری مواجه می‌شود که پیکسل‌های یکسان دارند اما برچسب‌های عمق متناقض؛ بهترین کاری که از آن برمی‌آید رگرسیون به میانگین توزیع α است و بخشی از خطا برای آن کاهش‌ناپذیر می‌شود. در مقابل، مدل پیشنهادی (دارای K) می‌تواند α را مستقیماً از مقدار فاصلهٔ کانونی بخواند و ابهام را به‌طور کامل رفع کند. به بیان دیگر، لرزش کانونی

[1] Focal Jitter

مسئلهٔ بازیابی مقیاس را برای مدل کنترل (فاقد K) به مسئله‌ای بدساخت[1] تبدیل می‌کند و بدین ترتیب فشار آموزشی صریحی برای بهره‌گیری از ورودی K ایجاد می‌نماید.

شکل ۶.۳ عدم تقارن این طراحی را میان دو بازوی مطالعه نشان می‌دهد.



شکل ۶.۳: منطق افزونه‌سازی لرزش کانونی و عدم تقارن عمدی آن میان دو بازوی مطالعهٔ کنترل‌شده. چون تبدیل $(f, Z) \mapsto (\alpha f, \alpha Z)$ تصویر را دقیقاً بدون تغییر باقی می‌گذارد (شکل ۱.۱-ب)، مدل کنترل (فاقد K) با پیکسل‌های یکسان و برچسب‌های عمق متناقض روبه‌رو می‌شود و بازیابی مقیاس برای آن به مسئله‌ای بدساخت تبدیل می‌گردد؛ در حالی که مدل پیشنهادی (دارای K) می‌تواند ضریب α را از ورودی K بخواند. این طراحی، اتکای مدل به پارامترهای دوربین را از یک گزینه به یک ضرورت آموزشی ارتقا می‌دهد.

دو نکتهٔ پیاده‌سازی اهمیت اساسی دارد. نخست، ماسک اعتبار پیکسل‌ها ویژگی حسگر و متعلق به عمق مقیاس‌نشده است؛ بنابراین آستانهٔ سقف عمق در ماسک نیز باید در α ضرب شود، وگرنه دقیقاً نقاط دوری که α در آن‌ها بیشترین اطلاعات را دارد از آموزش حذف می‌شوند. دوم، برای آنکه مقایسهٔ زوجی مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) کاملاً منصفانه باشد، سازوکار قفل بذر تصادفی[2] به زیرساخت آموزش افزوده شد که ترتیب داده‌ها، وارونه‌سازی‌های افقی، پنجره‌های زوم‌ـبرش و نمونه‌های α را میان دو اجرا یکسان می‌کند؛ صحت این سازوکار با برابری بیت‌به‌بیت مقدار تابع هزینه در گام صفر هر دو بازو راستی‌آزمایی شد. همچنین در آزمایش متناظر، ضریب λ در تابع هزینهٔ SiLog (بخش ۳-۸) از ۰٫۵ به ۰٫۳ کاهش یافت تا مؤلفهٔ خطای مقیاس سراسری — که دقیقاً همان اطلاعاتی است که K فراهم می‌کند — کمتر بخشوده شود.

[1] Ill-posed
[2] Seed Locking

۹.۷.۳ استفاده از تقطیر دانش

زیرساخت آموزش برای بررسی تقطیر دانش به‌عنوان یک قید نظارتی ضعیف توسعه یافته است. در این طرح اختیاری، مدل غیرمتریک Depth Anything V2 نقش معلم را دارد و پارامترهای `--use-distillation` و `--teacher-checkpoint` بارگذاری معلم را کنترل می‌کنند. این مسیر در مرحله‌های اکتشافی در نظر گرفته شد، اما جزئی از دستورالعمل نهایی تولید نتایج فصل چهارم نیست.

شایان ذکر است که در مجموعهٔ آزمایش‌های نهایی گزارش‌شده در فصل چهارم، به‌منظور جداسازی دقیق اثر پارامترهای درونی دوربین از سایر عوامل، سازوکار تقطیر دانش غیرفعال بوده و آموزش هر دو مدل (مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K)) صرفاً با نظارت مستقیم بر مبنای نقشه‌های عمق مرجع انجام شده است. قابلیت تقطیر دانش به‌عنوان بخشی از زیرساخت پیاده‌سازی حفظ شده و می‌تواند در پژوهش‌های آتی به‌کار گرفته شود.

۱۰.۷.۳ جمع‌بندی فرآیند آموزش

در مجموع، راهبرد نهایی بر فاین‌تیون نظارتی و کنترل‌شدهٔ یک مدل قدرتمند ازپیش‌آموزش‌یافته، فریزکردن انکودر، آموزش سر عمق و کدگذار دوربین، و زوج‌سازی کامل مدل پیشنهادی (دارای K) با مدل کنترل (فاقد K) استوار است. تقطیر دانش در اجرای نهایی غیرفعال ماند تا اثر پارامترهای دوربین از سایر سیگنال‌های آموزشی جدا شود.

در بخش بعدی، توابع هزینهٔ مورد استفاده در این فرآیند آموزش و نقش هر یک در هدایت مدل به‌سوی تخمین عمق متریک دقیق، به‌صورت رسمی و ریاضی تشریح خواهد شد.

۸.۳ توابع هزینه و معیارهای بهینه‌سازی

طراحی تابع هزینه در مسئلهٔ تخمین عمق تک‌دوربینه، نقش تعیین‌کننده‌ای در نوع اطلاعاتی دارد که مدل در فرآیند آموزش فرا می‌گیرد. از آنجا که هدف این پژوهش، دستیابی به تخمین عمق متریک پایدار و ناوابسته به تغییر پارامترهای درونی دوربین است، انتخاب تابع هزینه به‌گونه‌ای انجام شده که مدل بتواند مقیاس مطلق عمق را یاد بگیرد و در عین حال نسبت به تغییرات پارامترهای دوربین پایدار باشد.

در این پژوهش، بسته به نوع مدل (متریک یا نسبی)، از توابع هزینهٔ مناسبی استفاده شده است که در ادامه تشریح می‌شوند.

الگوریتم ۱.۳ روند آموزش مدل با پارامترهای هندسی

ورودی: مجموعه داده آموزش $D = \{(I_i, K_i, D_i^{GT})\}_{i=1}^{N}$، مدل پایهٔ Depth Anything V2 با وزن‌های θ_{enc} (فریز شده)، θ_{dec} (قابل آموزش)، θ_{cam} (شبکه کدگذار دوربین)
خروجی: مدل نهایی آموزش‌دیده

برای هر epoch از آموزش **انجام بده**
  برای هر دسته داده (batch) از D **انجام بده**
    تصویر $\leftarrow I$
    ماتریس درونی دوربین $\leftarrow K$
    نقشه عمق مرجع $\leftarrow D^{GT}$
    {پیش‌پردازش و کدگذاری پارامترهای دوربین}
    $\mathbf{k} \leftarrow$ نرمال‌سازی(K, W, H)
    $\mathbf{z}_k \leftarrow MLP(\mathbf{k}, \theta_{cam})$
    {استخراج ویژگی‌های بصری با تزریق توکن دوربین در انکودر}
    $F \leftarrow Encoder(I, \mathbf{z}_k, \theta_{enc})$
    {تخمین عمق}
    $D^{pred} \leftarrow Decoder(F, \theta_{dec})$
    {محاسبه تابع هزینه متریک}
    $\mathcal{L} \leftarrow \mathcal{L}_{SiLog}(D^{pred}, D^{GT})$
    {به‌روزرسانی وزن‌ها}
    $\theta_{dec}, \theta_{cam} \leftarrow Optimizer.step(\nabla\mathcal{L})$
  پایان حلقهٔ برای
پایان حلقهٔ برای

۱.۸.۳ هزینهٔ SiLog برای مدل‌های متریک

برای آموزش مدل متریک که هدف آن تولید تخمین عمق مطلق بر حسب متر است، از تابع هزینهٔ لگاریتمی تغییرناپذیر نسبت به مقیاس[1] استفاده می‌شود. این تابع هزینه که به اختصار SiLog نامیده می‌شود، به‌گونه‌ای طراحی شده که به‌جای تفاوت مستقیم مقادیر عمق، بر تفاوت لگاریتمی آن‌ها تمرکز دارد:

$$\mathcal{L}_{\text{SiLog}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2 - \frac{\lambda}{N^2}\left(\sum_{i=1}^{N} d_i\right)^2}, \tag{۱۶.۳}$$

که در آن N تعداد پیکسل‌های دارای برچسب معتبر، $d_i = \log D_i^{GT} - \log D_i^{pred}$ تفاوت لگاریتمی بین مقدار واقعی و مقدار پیش‌بینی‌شده در پیکسل i-ام است. ضریب λ معمولاً برابر ۰/۵ در نظر گرفته می‌شود تا

[1] Scale-Invariant Logarithmic Loss

تعادلی بین خطای مطلق و خطای نسبی برقرار گردد.

ویژگی مهم SiLog آن است که ضمن حفظ حساسیت به مقیاس مطلق عمق، باعث می‌شود خطا در فواصل دور و نزدیک به‌صورت متعادل‌تری در نظر گرفته شود. استفاده از فضای لگاریتمی، مدل را نسبت به خطاهای بزرگ در عمق‌های دور مقاوم‌تر می‌کند و از تمرکز بیش از حد بر خطاهای نقطه‌ای بسیار بزرگ جلوگیری می‌نماید.

۲.۸.۳ هزینهٔ تغییرناپذیر نسبت به مقیاس و جابجایی برای مدل‌های نسبی

برای آموزش مدل در حالت نسبی (basic mode) که خروجی آن دارای ابهام مقیاس است، از تابع هزینهٔ تغییرناپذیر نسبت به مقیاس و جابجایی[1] استفاده می‌شود. این تابع هزینه ابتدا پیش‌بینی مدل را با استفاده از روش کمترین مربعات بر مقدار واقعی منطبق می‌کند:

$$D^*_{pred} = s \cdot D_{pred} + t, \tag{۱۷.۳}$$

که در آن s و t مقیاس و جابجایی بهینه هستند که به‌صورت تحلیلی از طریق کمینه‌سازی خطای پیکسل‌به‌پیکسل محاسبه می‌شوند.

تابع هزینهٔ نهایی ترکیبی از خطای مطلق پیکسل‌ها ($L_۱$) و خطای تطابق گرادیان است:

$$\mathcal{L}_{\text{SSI}} = (۱ - \alpha)\mathcal{L}_{\text{pixel}} + \alpha\mathcal{L}_{\text{grad}}, \tag{۱۸.۳}$$

که در آن $\mathcal{L}_{\text{pixel}}$ خطای $L_۱$ میان تخمین منطبق‌شده و عمق واقعی است و $\mathcal{L}_{\text{grad}}$ خطای تطابق گرادیان در چند مقیاس را اندازه می‌گیرد. پارامتر α (معمولاً 0.5) وزن نسبی این دو مؤلفه را تعیین می‌کند. هزینهٔ گرادیان باعث می‌شود که لبه‌های عمق و ناپیوستگی‌ها در تصویر با دقت بیشتری بازسازی شوند.

[1] Scale-Shift Invariant Loss

۳.۸.۳ تمایز میان توابع هزینهٔ آموزش و معیارهای ارزیابی

نکته‌ای حائز اهمیت در این پژوهش، تمایز آگاهانه میان توابع هزینهٔ مورد استفاده در آموزش و معیارهای به‌کاررفته در ارزیابی مدل است. این تمایز از آن جهت ضروری است که اهداف آموزش و ارزیابی لزوماً یکسان نیستند.

۱.۳.۸.۳ توابع هزینه در آموزش

در مرحلهٔ آموزش:

- برای مدل‌های متریک از **SiLog** استفاده می‌شود
- برای مدل‌های نسبی از **Loss Invariant Scale-Shift** استفاده می‌شود

۲.۳.۸.۳ معیارهای ارزیابی

در مرحلهٔ ارزیابی، از معیارهای متنوع‌تری بسته به نوع مدل استفاده می‌شود:
برای مدل‌های متریک، از معیارهای استاندارد تخمین عمق استفاده می‌گردد:

- **خطای مطلق نسبی :(AbsRel)**

$$\text{AbsRel} = \frac{۱}{N}\sum_{i=۱}^{N}\frac{|D_i^{pred} - D_i^{GT}|}{D_i^{GT}} \tag{۱۹.۳}$$

- **RMSE:**

$$\text{RMSE} = \sqrt{\frac{۱}{N}\sum_{i=۱}^{N}(D_i^{pred} - D_i^{GT})^۲} \tag{۲۰.۳}$$

- **معیارهای دقت آستانه‌ای (δ_i):** درصد پیکسل‌هایی که $\max\left(\frac{D^{pred}}{D^{GT}}, \frac{D^{GT}}{D^{pred}}\right) < ۱/۲۵^i$ برای $i \in \{۱, ۲, ۳\}$

برای مدل‌های نسبی که دارای ابهام مقیاس هستند، ابتدا یک تراز مقیاس (معمولاً از طریق میانه) اعمال شده و سپس معیارهای فوق محاسبه می‌شوند. همچنین معیار SiLog که ذاتاً تغییرناپذیر نسبت به مقیاس است، برای مقایسهٔ این نوع مدل‌ها مناسب‌تر است.

جدول ۴.۳ این تمایزها را در یک نما جمع‌بندی می‌کند و نشان می‌دهد که چرا انتخاب نادرست فضای هدف — موضوع بخش ۶.۷.۳ — می‌تواند به شکست خاموش آموزش بینجامد.

جدول ۴.۳: جمع‌بندی توابع هزینه، فضای هدف آموزش و پروتکل ارزیابی متناظر برای دو حالت مدل. نکتهٔ کلیدی آن است که فضای هدف آموزش باید با فضای خروجی مدل از پیش‌آموزش‌دیده هم‌خوان باشد.

حالت مدل	فضای هدف آموزش	تابع هزینه	پروتکل ارزیابی و دلالت آن
نسبی (basic)	عمق معکوس $T_{\text{disp}} = ۱/\max(D^{GT}, d_{\min})$	تغییرناپذیر نسبت به مقیاس و جابجایی: $\mathcal{L}_{\text{SSI}} = (۱-\alpha)\mathcal{L}_{\text{pixel}} + \alpha\mathcal{L}_{\text{grad}}$	تراز مقیاس و جابجایی به‌ازای هر تصویر پیش از محاسبهٔ معیارها؛ این تراز، اطلاعات مقیاس مطلق — یعنی دقیقاً همان چیزی که K فراهم می‌کند — را حذف می‌کند
متریک (metric)	عمق مطلق بر حسب متر	لگاریتمی تغییرناپذیر نسبت به مقیاس: $\mathcal{L}_{\text{SiLog}}$ با ضریب λ	بدون هیچ ترازی؛ تنها پروتکلی که سهم اطلاعات هندسی دوربین در آن قابل مشاهده است

۴.۸.۳ نقش پارامترهای درونی دوربین در آموزش

نقش ماتریس مشخصهٔ دوربین در این چارچوب، نه از طریق افزودن یک مؤلفهٔ مستقیم به تابع هزینه، بلکه از طریق تزریق اطلاعات هندسی به معماری مدل ایفا می‌شود. با در اختیار داشتن اطلاعات صریح پارامترهای درونی، مدل قادر است تخمین‌های متریک خود را با هندسهٔ واقعی تصویربرداری تطبیق دهد.

در مدل پیشنهادی (دارای K)، پارامترهای درونی دوربین از طریق یک واحد کدگذار مجزا پردازش شده و به‌صورت توکن دوربین به لایه‌های میانی Transformer Vision تزریق می‌شوند. این اطلاعات به مدل امکان می‌دهد تا در فرآیند یادگیری، رابطهٔ میان ویژگی‌های تصویری و مقیاس واقعی عمق را با در نظر گرفتن مشخصات هندسی دوربین یاد بگیرد.

به این ترتیب، تابع هزینهٔ SiLog که ذاتاً به مقیاس مطلق عمق حساس است، می‌تواند به‌درستی مدل را برای

تولید تخمین‌های متریک پایدار هدایت کند. این پایداری در مواجهه با داده‌هایی که تنها در پارامترهای درونی دوربین تفاوت دارند، در آزمایش‌های تجربی به‌وضوح مشاهده می‌شود.

۵.۸.۳ حساسیت RMSE به پارامترهای دوربین و کاربرد آن در ارزیابی

معیار RMSE که در این پژوهش برای ارزیابی مدل‌های متریک استفاده شده، به‌صورت مستقیم به مقیاس عمق وابسته است:

$$\mathrm{RMSE} = \sqrt{\frac{۱}{N}\sum_{i=۱}^{N}(D_i^{pred} - D_i^{GT})^۲}. \tag{۲۱.۳}$$

این وابستگی باعث می‌شود RMSE نسبت به تغییر پارامترهای درونی دوربین، به‌ویژه فاصلهٔ کانونی، حساس باشد. این ویژگی، هرچند در نگاه اول یک محدودیت به نظر می‌رسد، اما در واقع ابزار مناسبی برای آشکارسازی ناپایداری متریک مدل‌های عمق محسوب می‌شود.

در این پژوهش، از این حساسیت به‌عنوان معیاری برای سنجش پایداری مدل استفاده شده است. مدلی که نسبت به تغییر پارامترهای دوربین پایدار باشد، باید تغییرات کمی در RMSE از خود نشان دهد، حتی زمانی که تصاویر با مشخصات هندسی متفاوت تصویربرداری شده باشند.

۶.۸.۳ جمع‌بندی

در این بخش، توابع هزینهٔ مورد استفاده در آموزش مدل‌های متریک و نسبی تشریح شد. استفاده از SiLog برای مدل‌های متریک، ضمن حفظ حساسیت به مقیاس مطلق عمق، مدل را نسبت به خطاهای بزرگ مقاوم می‌کند. تزریق پارامترهای درونی دوربین به معماری مدل، امکان یادگیری رابطهٔ صحیح میان ویژگی‌های تصویری و عمق واقعی را فراهم می‌آورد. در فصل بعد، نتایج تجربی حاصل از این طراحی و تحلیل آماری پایداری مدل نسبت به تغییر مشخصات دوربین ارائه خواهد شد.

۹.۳ جزئیات پیاده‌سازی

در این بخش، جزئیات فنی مربوط به پیاده‌سازی مدل پیشنهادی (دارای K)، تنظیمات آموزش، پیش‌پردازش داده‌ها، ساختار کد و ملاحظات عملی لازم برای بازتولید نتایج ارائه می‌شود. هدف از این بخش، مستندسازی دقیق تمامی تصمیمات پیاده‌سازی است تا امکان تکرار آزمایش‌ها و تحلیل نتایج برای پژوهشگران دیگر فراهم شود.

۱.۹.۳ چارچوب نرم‌افزاری و زیرساخت محاسباتی

تمامی پیاده‌سازی‌ها در چارچوب PyTorch (نسخهٔ 1.12 و بالاتر) انجام شده و کد پایه بر اساس مخزن رسمی Depth Anything V2 توسعه داده شده است. معماری کلی شامل یک انکودر مبتنی بر DINOv2 Vision Transformer و یک دیکودر از خانوادهٔ Dense Prediction Transformer (DPT) است.

آموزش مدل‌ها بر روی پردازنده‌های گرافیکی NVIDIA با معماری Ampere یا جدیدتر انجام شده و تمامی آزمایش‌ها به‌صورت تک‌کارت گرافیک و با دقت ممیز شناور ۳۲ بیتی اجرا شده‌اند. برای تضمین تکرارپذیری، بذر تصادفی[1] برای کتابخانه‌های PyTorch، NumPy و Python در ابتدای هر آزمایش ثابت (معمولاً بر روی مقدار ۴۲) تنظیم شده است.

۲.۹.۳ پیش‌پردازش داده‌ها

تمامی تصاویر ورودی پیش از ورود به شبکه، به‌صورت یکنواخت پیش‌پردازش می‌شوند. در تنظیم پایه، تصاویر با حفظ نسبت ابعاد به کران پایین ۵۱۸ پیکسل تغییر اندازه داده شده و سپس برش‌های تصادفی ۵۱۸×۵۱۸ برای یکسان‌سازی ابعاد دسته در مرحلهٔ آموزش اعمال می‌گردد؛ راهبردی مطابق با تنظیمات مدل پایهٔ Depth Anything V2 [۱۳]. در آزمایش‌های نهایی مبتنی بر افزونه‌سازی هندسی (بخش ۷.۷.۳)، این برش تصادفی سادهٔ ناآگاه از K کنار گذاشته شده و کل زنجیرهٔ برش و تغییر اندازه به‌صورت آگاه از پارامترهای دوربین و با به‌روزرسانی دقیق ماتریس K انجام می‌شود؛ در این حالت، چارچوب ورودی شبکه به‌صورت ثابت ۶۸۶ × ۵۱۸ (قاب کامل ۳:۴ در کران پایین ۵۱۸ و مضرب ۱۴) انتخاب شده است.

مقادیر پیکسل تصاویر به بازهٔ [۰, ۱] نرمال‌سازی شده و سپس با استفاده از میانگین و واریانس داده‌های ImageNet استانداردسازی می‌شوند:

[1] Random Seed

$$I_{norm} = \frac{I - \mu_{ImageNet}}{\sigma_{ImageNet}}, \tag{۲۲.۳}$$

که در آن $\mu_{ImageNet} = [۰/۴۸۵, ۰/۴۵۶, ۰/۴۰۶]$ و $\sigma_{ImageNet} = [۰/۲۲۹, ۰/۲۲۴, ۰/۲۲۵]$ هستند. این مرحله به‌ویژه برای سازگاری با وزن‌های ازپیش‌آموزش‌دیدهٔ انکودر DINOv2 ضروری است.

۳.۹.۳ مدیریت و نرمال‌سازی پارامترهای درونی دوربین

برای هر تصویر ورودی، ماتریس مشخصهٔ دوربین K استخراج شده و پارامترهای اصلی آن شامل فاصلهٔ کانونی در راستاهای افقی و عمودی (f_x, f_y) و مختصات نقطهٔ اصلی (c_x, c_y) مورد استفاده قرار می‌گیرند.

به‌منظور جلوگیری از ناپایداری عددی و تسهیل فرآیند یادگیری، این پارامترها پیش از ورود به شبکه نرمال‌سازی می‌شوند:

$$\tilde{f}_x = \frac{f_x}{W}, \quad \tilde{f}_y = \frac{f_y}{H}, \quad \tilde{c}_x = \frac{c_x}{W}, \quad \tilde{c}_y = \frac{c_y}{H}, \tag{۲۳.۳}$$

که در آن W و H به‌ترتیب عرض و ارتفاع تصویر ورودی هستند.

بردار نهایی پارامترهای دوربین به‌صورت

$$\mathbf{k} = [\tilde{f}_x, \tilde{f}_y, \tilde{c}_x, \tilde{c}_y] \tag{۲۴.۳}$$

تعریف شده و همراه با مؤلفه‌های میدان دید و نسبت ابعاد (مطابق بخش ۳-۶)، یک بردار توصیف‌گر نه‌بعدی را تشکیل می‌دهد. این بردار از طریق یک شبکهٔ تمام‌متصل دولایه (با تابع فعال‌سازی GELU) به فضایی هم‌بعد با بازنمایی نهان انکودر (برای ViT-L برابر ۱۰۲۴) نگاشت شده و سپس توسط چهار بلوک ترنسفورمر پالایش می‌شود. خروجی این واحد، پس از اعمال دروازهٔ خروجی صفرمقدار اولیه، به‌عنوان توکن دوربین $\mathbf{z}_k$ به دنبالهٔ توکن‌های انکودر تزریق می‌گردد.

۴.۹.۳ تنظیمات آموزش و راهبرد بهینه‌سازی

مدل پیشنهادی (دارای K) با استفاده از بهینه‌ساز AdamW [۲۸] آموزش داده شده است. نرخ یادگیری پایه برای انکودر DINOv2 برابر با ۵×۱۰^{-۶} در نظر گرفته شده، در حالی که برای دیکودر و لایه‌های افزوده‌شده (شامل کدگذار دوربین و سر تخمین عمق)، از یک ضریب تقویتی (معمولاً ۱۰ برابر) استفاده شده که نرخ یادگیری این بخش‌ها را به ۵×۱۰^{-۵} می‌رساند.

پارامترهای بهینه‌ساز به‌صورت زیر تنظیم شده‌اند:

- ضرایب ممنتوم: $\beta_۱ = ۰/۹$، $\beta_۲ = ۰/۹۹۹$
- ضریب کاهش وزن (Weight Decay): ۰/۰۱
- پارامتر ϵ برای پایداری عددی: $۱۰^{-۸}$

در راستای جلوگیری از بیش‌برازش و حفظ دانش ازپیش‌آموخته‌شده، وزن‌های انکودر DINOv2 در طول آموزش فریز شده‌اند. این تنظیم، مطابق با گزینهٔ پیش‌فرض `--freeze-dinov2` در اسکریپت آموزش اعمال شده است. فریز کردن انکودر باعث می‌شود تنها لایه‌های سر شبکه[1] و واحد کدگذار دوربین آموزش ببینند، که این امر از واگرایی نمایش‌های ویژگی جلوگیری کرده و همگرایی سریع‌تر را تضمین می‌کند.

برای زمان‌بندی نرخ یادگیری، از یک استراتژی ترکیبی شامل warmup خطی در ابتدای آموزش و سپس کاهش چندجمله‌ای (Polynomial Decay) با توان ۰/۹ استفاده شده است. تعداد دوره‌های آموزش معمولاً بین ۲۰ تا ۴۰ epoch بوده و اندازهٔ دسته بسته به حافظهٔ موجود بین ۲ تا ۸ تنظیم شده است. در صورت محدودیت حافظه، از تجمع گرادیان[2] برای افزایش اندازهٔ دستهٔ مؤثر استفاده شده است.

جدول ۵.۳ تمامی ابرپارامترهای آموزش را در یک نما گرد آورده است. تأکید می‌شود که این مقادیر برای هر دو بازوی مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) کاملاً یکسان‌اند؛ تنها تفاوت دو بازو، وجود یا نبود ورودی K است.

۵.۹.۳ دیتاست‌های مورد استفاده در پیاده‌سازی

پیاده‌سازی مدل بر روی مجموعه‌ای متنوع از دیتاست‌های داخلی و خارجی انجام شده است که شامل صحنه‌های indoor و outdoor با مشخصات تصویربرداری متفاوت هستند. از جمله دیتاست‌های مورد استفاده

[1] Head Layers
[2] Gradient Accumulation

جدول ۵.۳: ابرپارامترهای آموزش مطالعهٔ کنترل‌شدهٔ نهایی. این مقادیر برای بازوی مدل پیشنهادی (دارای K) و بازوی مدل کنترل (فاقد K) یکسان‌اند؛ تنها متغیر مورد مطالعه، دسترسی مدل به ورودی K است.

مؤلفه	مقدار	توضیح
بهینه‌ساز	AdamW [۲۸]	$\beta_1 = ۰/۹$، $\beta_2 = ۰/۹۹۹$، $\epsilon = ۱۰^{-۸}$
ضریب کاهش وزن	۰/۰۱	—
نرخ یادگیری پایه	۵×۱۰^{-۶}	برای انکودر DINOv2
ضریب تقویت نرخ یادگیری سر شبکه	۱۰	نرخ مؤثر دیکودر و کدگذار دوربین: ۵×۱۰^{-۵}
زمان‌بندی نرخ یادگیری	گرم‌سازی خطی + کاهش چندجمله‌ای	توان کاهش برابر ۰/۹
انکودر	ViT-L، فریزشده	گزینهٔ `--freeze-dinov2`
تعداد دوره‌ها	۴۰	—
اندازهٔ دسته	۲، با تجمع گرادیان ۲	اندازهٔ دستهٔ مؤثر: ۴
برش گرادیان	۱/۰	برای پایداری عددی
چارچوب ورودی شبکه	۵۱۸×۶۸۶	قاب کامل ۳:۴، مضرب اندازهٔ وصله
دقت عددی	ممیز شناور ۳۲ بیتی	تک‌کارت گرافیک
بذر تصادفی	۴۲، قفل‌شده میان دو بازو	راستی‌آزمایی با برابری بیت‌به‌بیت هزینهٔ گام صفر
افزونه‌سازی هندسی	برش ۳:۴ با به‌روزرسانی دقیق K	شکل ۵.۳
لرزش کانونی	$m = ۱/۶$	شکل ۶.۳
ضریب λ در SiLog	۰/۵ (و ۰/۳ در دور سوم)	مطالعهٔ فروکاست بخش ۸.۵.۴
تقطیر دانش	غیرفعال	برای جداسازی اثر K از سیگنال‌های آموزشی جانبی

می‌توان به موارد زیر اشاره کرد:

- **VKITTI2:** دیتاست شبیه‌سازی‌شده برای صحنه‌های رانندگی با truth ground دقیق عمق و پارامترهای دوربین
- **KITTI:** دیتاست واقعی رانندگی خودکار با اطلاعات عمق از لایدار
- **NYUv2:** دیتاست صحنه‌های داخلی با عمق حاصل از سنسور Kinect
- **Hypersim:** دیتاست سنتتیک با تنوع بالا در صحنه‌های داخلی و کنترل کامل بر پارامترهای دوربین

این تنوع داده در مرحلهٔ اکتشافی برای سنجش پایداری بین‌دیتاستی به‌کار رفت. قالب دقیق فهرست و کالیبراسیون به مجموعه‌داده وابسته است: پارامترها ممکن است در فایل npy.، فایل متنی کالیبراسیون یا مستندات

رسمی مجموعه‌داده موجود باشند. مطالعهٔ کنترل‌شدهٔ نهایی، همان‌گونه که پیش‌تر بیان شد، تنها از تقسیم بدون نشت NYUv2 استفاده می‌کند.

۶.۹.۳ ساختار کد و اسکریپت‌های اصلی

ساختار کد به‌گونه‌ای طراحی شده که اجزای اصلی شامل ماژول‌های بارگذاری داده، تعریف مدل، آموزش و ارزیابی به‌صورت مجزا پیاده‌سازی شوند. ساختار کلی پروژه به شرح زیر است:

```
project_root/
|-- models/
|   |-- raw_models/
|   |   |-- DepthAnythingV2/
|   |   |-- DepthAnythingV2-revised/
|   |   `-- Depth-Anything-3/
|   `-- base.py
|-- datasets/
|   |-- base.py
|   `-- training_datasets.py
|-- src/
|   |-- pipeline.py
|   |-- metrics.py
|   `-- visualization.py
|-- train.py
|-- evaluate.py
`-- compare_models.py
```

اسکریپت `train.py` مسئول اجرای فرآیند آموزش بوده و از طریق آرگومان‌های ورودی، امکان تنظیم پارامترهای زیر را فراهم می‌کند:

- `--use-camera-intrinsics`: فعال‌سازی استفاده از پارامترهای دوربین

- `--model-type`: انتخاب نوع مدل (metric یا basic)
- `--encoder`: انتخاب اندازهٔ انکودر (vits, vitb, vitl, vitg)
- `--use-distillation`: فعال‌سازی تقطیر دانش
- `--freeze-dinov2`: فریز کردن وزن‌های انکودر
- `--focal-jitter` و `--silog-lambda`: تنظیم لرزش کانونی و ضریب مقیاس در هزینهٔ SiLog
- `--seed`: قفل‌کردن ترتیب داده و جریان‌های تصادفی برای زوج‌های کنترل‌شده
- `--accumulate-grad`: تجمع گرادیان برای دستیابی به اندازهٔ دستهٔ مؤثر بزرگ‌تر

علاوه بر اسکریپت آموزش، ابزارهای اصلی به شرح زیر هستند:

- `evaluate_paper.py` برای بازتولید پروتکل ارزیابی مقاله؛
- `tools/paired_stats.py` برای آزمون زوجی، ویلکاکسون و بوت‌استرپ ده‌هزارباره؛
- `tools/k_sweep.py` و `tools/k_knockout.py` برای آزمون‌های سازوکاری و علّی؛
- `tools/eval_fov_varied.py` برای سنجش تعمیم به میدان دیدهای متفاوت.

صحت تبدیل لرزش کانونی، منطق ماسک اعتبار و اعتبارسنجی آرگومان‌ها نیز با ده آزمون واحد در فایل `tests/test_focal_jitter.py` پوشش داده شده است.

دستورالعمل مدل نهاییِ فروکاست — مدلی که قوی‌ترین شاهد استفاده از K را با $\lambda = ۰/۵$ ارائه می‌کند — به‌صورت زیر است. برای بازوی مدل کنترل (فاقد K)، دو گزینهٔ پایانی با `--model-variant base` جایگزین می‌شوند؛ سایر آرگومان‌ها بدون تغییر می‌مانند.

```
python train.py --encoder vitl --datasets nyu --model-type metric \
  --min-depth 0.001 --max-depth 20 --epochs 40 --bs 2 \
  --accumulate-grad 2 --lr 5e-6 --freeze-dinov2 \
  --head-lr-multiplier 10 --grad-clip 1 --focal-jitter 1.6 \
  --silog-lambda 0.5 --seed 42 --pretrained-from <basic_vitl.pth> \
  --save-path <output_dir> --model-variant revised \
  --use-camera-intrinsics
```

این مدل با AbsRel برابر ۰٫۰۸۳۴ و شیب جاروی K برابر ۰٫۰۳ ± ۰٫۹۷، بهترین مدل دارای لرزش کانونی است. اگر تنها کمینه‌کردن خطای خام روی هندسهٔ استاندارد NYUv2 هدف باشد، مدل متریک دور نخست بدون لرزش کانونی با AbsRel برابر ۰٫۰۷۹۱ بهتر است؛ اما برای ادعای اصلی پایان‌نامه دربارهٔ مصرف علّی K، مدل فروکاست نهایی انتخاب مناسب‌تری است.

۷.۹.۳ ملاحظات عملی و محدودیت‌ها

در پیاده‌سازی عملی، برخی محدودیت‌ها و چالش‌ها شناسایی و مدیریت شده‌اند:

- **محدودیت حافظهٔ GPU:** با توجه به حجم بالای مدل‌های Vision، Transformer از تجمع گرادیان استفاده شده است. به‌طور مشخص، آموزش ViT-L در چارچوب ورودی ۶۸۶ × ۵۱۸ با اندازهٔ دستهٔ ۴ از حافظهٔ ۲۴ گیگابایتی کارت گرافیک فراتر می‌رفت؛ از این رو، اندازهٔ دستهٔ ۲ همراه با تجمع گرادیان با گام ۲ (اندازهٔ دستهٔ مؤثر ۴) به‌کار گرفته شد.

- **عدم دسترسی به پارامترهای دقیق دوربین:** برای برخی دیتاست‌ها، پارامترهای دوربین از روی اطلاعات EXIF تصاویر یا مستندات دیتاست استخراج یا تخمین زده شده است. در مواردی که اطلاعات دقیق در دسترس نبوده، از مقادیر متعارف (مانند $f = W$ برای تصاویر با FoV استاندارد) استفاده شده است.

- **تنوع در کیفیت ground truth:** دیتاست‌های مختلف دارای سطوح متفاوتی از دقت در نقشه‌های عمق مرجع هستند. برای رفع این مسئله، در فرآیند ارزیابی از masking برای حذف پیکسل‌های نامعتبر استفاده شده است. در ارزیابی استاندارد NYUv2، مطابق پروتکل مرجع، از نقشه‌های عمق خام ثبت‌شده[1] (و نه نسخه‌های درون‌یابی‌شده) به‌عنوان مرجع استفاده شده است.

- **تکرارپذیری و استنباط آماری:** علی‌رغم ثابت نگه‌داشتن بذرهای تصادفی، برخی عملیات GPU ذاتاً غیرقطعی هستند. برای استنباط آماری معتبر، به‌جای اتکا به مقایسهٔ میانگین‌های کل، از آزمون‌های زوجی در سطح تک‌تصویر (آزمون t زوجی و آزمون علامت‌دار ویلکاکسون بر روی ۶۵۴ تصویر آزمون) به‌همراه بازه‌های اطمینان بوت‌استرپ استفاده شده است. آموزش با چند بذر مستقل تکرار نشده است؛ ازاین‌رو، برآورد و گزارش انحراف معیار میان اجراها یکی از محدودیت‌ها و پیشنهادهای پژوهش آینده است.

[1] Raw Depth

۸.۹.۳ اشکال‌زدایی و درس‌های پیاده‌سازی

در مسیر رسیدن به پیکربندی نهایی، سه اشکال اساسی در زنجیرهٔ آموزش و ارزیابی شناسایی، ریشه‌یابی و برطرف شد که مستندسازی آن‌ها هم برای تکرارپذیری پژوهش و هم به‌عنوان درس‌های روش‌شناختی حائز اهمیت است:

۱. **ناهم‌خوانی فضای هدف و فروریزش خاموش سر شبکه:** همان‌گونه که در بخش ۶.۷.۳ تشریح شد، آموزش مدل نسبی با هدف عمق مستقیم، به فروریزش خروجی به مقدار ثابت صفر منجر می‌شد؛ شکستی که در منحنی هزینهٔ آموزش قابل مشاهده نبود و تنها از طریق پایش معیارهای اعتبارسنجی (که در مقداری ثابت منجمد می‌شدند) و سپس بازتولید کنترل‌شدهٔ گام‌به‌گام با ثبت آماره‌های خروجی و گرادیان در هر گام، ریشه‌یابی شد. راه‌حل، انتقال هدف آموزش به فضای عمق معکوس بود.

۲. **به‌روزنشدن ماتریس K در افزونه‌سازی داده:** برش تصادفی متداول، نقطهٔ اصلی مؤثر تصویر را تغییر می‌داد بدون آنکه K اصلاح شود؛ در نتیجه کدگذار دوربین ورودی هندسی نادرست دریافت می‌کرد. راه‌حل، طراحی زنجیرهٔ افزونه‌سازی هندسی دقیق (بخش ۷.۷.۳) بود که صحت آن با راستی‌آزمایی عددی میدان دید نمونه‌های تولیدی تأیید شد.

۳. **ناهم‌خوانی معماری سر شبکه در ارزیابی مدل‌های متریک:** نقاط وارسی متریک آموزش‌دیده در این پژوهش دارای سر خروجی مبتنی بر ReLU هستند، در حالی که بارگذار ارزیابی، به‌طور پیش‌فرض معماری سر سیگموئیدی ضرب‌در-عمق-بیشینهٔ نقاط وارسی رسمی را می‌ساخت. از آنجا که نام و ابعاد پارامترها در هر دو معماری یکسان است، بارگذاری وزن‌ها بدون هیچ خطایی انجام می‌شد اما خروجی عمق حدود چهار برابر بیش‌برآورد می‌گردید (AbsRel ظاهری حدود ۲٫۹۷ در برابر مقدار صحیح نزدیک ۰٫۰۸). راه‌حل، ثبت نشانگر `from_train_py` در فرادادهٔ نقطهٔ وارسی و انتخاب خودکار معماری صحیح در زمان ارزیابی بود.

جدول ۶.۳ این سه مورد را در قالب زنجیرهٔ «نشانه، علت ریشه‌ای، راه‌حل» جمع‌بندی می‌کند.

درس مشترک هر سه مورد آن است که در زنجیره‌های یادگیری عمیق، شکست‌ها می‌توانند کاملاً خاموش باشند: منحنی هزینه طبیعی به نظر می‌رسد، بارگذاری وزن‌ها بدون خطا انجام می‌شود، و تنها راستی‌آزمایی‌های مستقل انتها-به-انتها (بازتولید نتایج مرجع منتشرشده، پایش انجماد معیارهای اعتبارسنجی، و بررسی آماره‌های خروجی مدل) قادر به آشکارسازی آن‌ها هستند. بر همین اساس، پیش از اجرای آزمایش‌های اصلی، صحت کل زنجیرهٔ ارزیابی این پژوهش از طریق بازتولید نتایج گزارش‌شدهٔ مقالهٔ Depth Anything V2 روی پروتکل استاندارد راستی‌آزمایی شد (فصل چهارم).

جدول ۶.۳: سه اشکال خاموش شناسایی و رفع‌شده در زنجیرهٔ آموزش و ارزیابی. وجه مشترک هر سه، آن است که هیچ‌یک پیام خطایی تولید نمی‌کردند و تنها با راستی‌آزمایی مستقل انتها-به-انتها آشکار شدند.

اشکال	نشانهٔ ظاهری	علت ریشه‌ای	راه‌حل اعمال‌شده
نشت تقسیم‌بندی NYUv2	خطای اعتبارسنجی به‌شکل غیرعادی پایین؛ نتایج خوش‌بینانه اما غیرقابل مقایسه با ادبیات	تقسیم تصادفی روی کل ۱۴۴۹ تصویر برچسب‌دار، که ۵۲۷ تصویر از ۶۵۴ تصویر آزمون استاندارد را وارد آموزش می‌کرد	حذف کامل تصاویر تقسیم آزمون Eigen پیش از ساخت مجموعه‌های آموزش و اعتبارسنجی
ناهم‌خوانی فضای هدف آموزش	منحنی هزینه کاملاً طبیعی، اما معیار اعتبارسنجی در مقداری ثابت منجمد و خروجی شبکه صفر	هدف آموزش در فضای عمق مستقیم، در حالی که مدل از پیش‌آموزش‌دیده در فضای عمق معکوس آموزش دیده است؛ در تعامل با محدودسازی ضریب مقیاس، بهینهٔ عملی تابع هزینه خروجی ثابت صفر می‌شود	انتقال هدف مدل‌های نسبی به فضای عمق معکوس و ثبت نوع فضای هدف در فرادادهٔ هر نقطهٔ وارسی
ناهم‌خوانی معماری سر شبکه در ارزیابی	بارگذاری وزن‌ها بدون هیچ خطا، اما AbsRel ظاهری حدود ۲٫۹۷ به‌جای مقدار صحیح نزدیک ۰٫۰۸	سر خروجی نقاط وارسی این پژوهش مبتنی بر ReLU است، در حالی که بارگذار ارزیابی سر سیگموئیدی نقاط وارسی رسمی را می‌ساخت؛ نام و ابعاد پارامترها در هر دو یکسان است	ثبت نشانگر `from_train_py` در فراداده و انتخاب خودکار معماری صحیح در زمان ارزیابی

۹.۹.۳ جمع‌بندی

در این بخش، جزئیات پیاده‌سازی مدل پیشنهادی (دارای K) شامل چارچوب نرم‌افزاری، پیش‌پردازش داده‌ها، مدیریت پارامترهای درونی دوربین، تنظیمات آموزش، دیتاست‌های مورد استفاده و ساختار کد تشریح شد. این توضیحات، مبنای فنی لازم برای درک و تحلیل نتایج تجربی ارائه‌شده در فصل بعد را فراهم می‌کنند و امکان بازتولید آزمایش‌ها توسط سایر پژوهشگران را تسهیل می‌نمایند.

۱۰.۳ جمع‌بندی فصل

در این فصل، چارچوب روش‌شناختی پژوهش حاضر برای استفاده از یک مدل بنیادین تخمین عمق تک‌دوربینه و فاین‌تیون نظارتی آن با هدف دستیابی به عمق متریک پایدار ارائه شد. تمرکز اصلی فصل بر تحلیل نقش هندسهٔ تصویربرداری و به‌ویژه ماتریس مشخصهٔ دوربین در فرآیند یادگیری عمق، و ارائهٔ یک راهکار عملی برای ادغام صریح این اطلاعات در مدل بود.

در ابتدا، مسئلهٔ تخمین عمق به‌صورت یک نگاشت تابعی از تصویر ورودی و پارامترهای درونی دوربین صورت‌بندی شد و تفاوت مفهومی میان عمق غیرمتریک و عمق متریک مورد بررسی قرار گرفت. نشان داده

شد که در غیاب اطلاعات هندسی دوربین، مدل‌های تک‌دوربینه—حتی در صورت عملکرد مناسب از منظر معیارهای نرمال‌شده—قادر به تولید تخمین‌هایی با مقیاس فیزیکی پایدار نیستند.

سپس، با تکیه بر مدل هندسی سوراخ‌سوزنی، نقش پارامترهای فاصلهٔ کانونی و نقطهٔ اصلی تصویر در تبدیل مختصات تصویری به عمق متریک تشریح شد. این تحلیل هندسی، مبنای نظری لازم برای انگیزه‌بخشی به روش پیشنهادی را فراهم کرد و روشن ساخت که نادیده گرفتن ماتریس مشخصهٔ دوربین می‌تواند به ناپایداری قابل‌توجه در معیارهای حساس به مقیاس، به‌ویژه RMSE، منجر شود.

در ادامه، معماری پایهٔ Depth Anything V2 به‌عنوان یک مدل بنیادین قدرتمند در تخمین عمق تک‌تصویری معرفی شد. این معماری، با وجود تعمیم‌پذیری بالا در سناریوهای صفرنمونه، ذاتاً برای تولید عمق غیرمتریک طراحی شده است. بر همین اساس، روش پیشنهادی این پژوهش بر افزودن یک مسیر صریح برای پردازش پارامترهای درونی دوربین و ادغام آن‌ها به‌صورت توکن دوربین در دنبالهٔ توکن‌های انکودر متمرکز شد، به‌گونه‌ای که اطلاعات هندسی از اطلاعات صرفاً تصویری تفکیک شده و از طریق سازوکار توجه، به‌صورت کنترل‌شده در فرآیند استنتاج عمق وارد شوند.

راهبرد پیشنهادی شامل نرمال‌سازی پارامترهای ماتریس مشخصهٔ دوربین، نگاشت آن‌ها به یک توکن نهان، و تزریق دروازه‌بانی‌شدهٔ این بازنمایی هندسی به انکودر مدل بود. این طراحی، به‌واسطهٔ دروازهٔ خروجی صفرمقدار اولیه، بدون ایجاد اختلال در رفتار اولیهٔ انکودر ازپیش‌آموزش‌دیده، امکان یادگیری تدریجی وابستگی‌های متریک را فراهم می‌کند. در آزمایش‌های نهایی، آموزش با نگاشت عمق مرجع و بدون تقطیر انجام شد تا سهم خالص K در مقایسه با مدل کنترل (فاقد K) هم‌داده قابل اندازه‌گیری باشد.

در بخش‌های پایانی فصل، توابع هزینه، معیارهای ارزیابی و جزئیات پیاده‌سازی به‌صورت کامل تشریح شدند. تمایز میان معیارهای نرمال‌شده مانند AbsRel و SiLog، در مقابل معیارهای حساس به واحد فیزیکی مانند RMSE، به‌عنوان یک نکتهٔ کلیدی در تحلیل عملکرد مدل‌های متریک برجسته شد. این تمایز، چارچوب مفهومی لازم برای تفسیر نتایج تجربی را فراهم می‌کند [۲، ۱۴].

در مجموع، این فصل یک چارچوب منسجم نظری و عملی برای ادغام اطلاعات درونی دوربین در تخمین عمق تک‌دوربینه ارائه داد. فصل بعد هم نتایج اکتشافی بین‌دیتاستی و هم مطالعهٔ نهایی کنترل‌شدهٔ NYUv2 را گزارش می‌کند و تأثیر افزودن ماتریس مشخصهٔ دوربین بر مقیاس متریک، دقت عددی و رفتار علّی مدل را تحلیل می‌نماید.

فصل ۴

نتایج

۱.۴ مقدمه

این فصل به ارائه، تحلیل و تفسیر نتایج تجربی حاصل از پیاده‌سازی و ارزیابی روش پیشنهادی در این پژوهش اختصاص دارد. در حالی‌که فصل سوم به تشریح چارچوب نظری، معماری مدل و راهبردهای آموزش پرداخته است، تمرکز فصل حاضر بر پاسخ‌گویی عملی و داده‌محور به سوالات تحقیق و بررسی میزان تحقق اهداف پژوهش از طریق آزمایش‌های کنترل‌شده و قابل تکرار است. بدین منظور، نتایج حاصل از اجرای مدل‌ها بر روی مجموعه‌ای متنوع از دیتاست‌های استاندارد ارائه شده و با استفاده از معیارهای کمی و تحلیل‌های آماری مورد ارزیابی قرار می‌گیرند.

هدف اصلی این پژوهش، ارتقای مدل Depth Anything V2 از یک مدل تخمین عمق غیرمتریک[1] به مدلی با خروجی عمق متریک پایدار[2] و کاهش حساسیت آن نسبت به تغییر پارامترهای دوربین، به‌ویژه ماتریس مشخصهٔ دوربین[3] بوده است. ازاین‌رو، در این فصل به‌طور خاص بررسی می‌شود که آیا ادغام صریح اطلاعات هندسی دوربین در فرآیند تخمین عمق، می‌تواند ناپایداری مقیاس و وابستگی به نوع دوربین را که در مدل‌های تک‌دوربینهٔ مرسوم مشاهده می‌شود، کاهش دهد یا خیر.

برای ارزیابی جامع عملکرد مدل پیشنهادی (دارای K)، مجموعه‌ای از آزمایش‌های کمی بر روی دیتاست‌های متنوع شامل محیط‌های داخلی، خارجی، واقعی و مصنوعی انجام شده است. در این آزمایش‌ها، مدل پیشنهادی

[1] Non-metric Depth
[2] Metric Depth
[3] Camera Intrinsics

(دارای K) با مدل پایهٔ Depth Anything V2 مقایسه شده و عملکرد آن با استفاده از معیارهای متداول ارزیابی عمق، از جمله معیارهای غیرمتریک و متریک، سنجیده می‌شود. همچنین، به‌منظور بررسی پایداری نتایج، از آزمون‌های آماری و بازه‌های اطمینان استفاده شده تا معناداری تفاوت‌های مشاهده‌شده به‌صورت دقیق تحلیل گردد. نتایج ارائه‌شده در این فصل، مبنای قضاوت نهایی در خصوص کارایی و اعتبار روش پیشنهادی را فراهم می‌کنند.

۲.۴ سوالات تحقیق و فرضیات تجربی

در فصل نخست این پژوهش، مجموعه‌ای از سوالات تحقیق با هدف بررسی امکان دستیابی به عمق متریک پایدار در چارچوب تخمین عمق تک‌دوربینه و نقش اطلاعات هندسی دوربین در این فرآیند مطرح شد. در این فصل، این سوالات بر اساس نتایج تجربی به‌دست‌آمده مورد ارزیابی و پاسخ‌گویی قرار می‌گیرند.

۱.۲.۴ بازگویی سوالات تحقیق

سوالات اصلی تحقیق به‌صورت زیر قابل بازبیان هستند:

- آیا مدل پایهٔ Depth Anything V2، قادر به تولید عمق متریک پایدار در دیتاست‌هایی با دوربین‌ها و پارامترهای تصویربرداری متفاوت است؟
- آیا ادغام صریح ماتریس مشخصهٔ دوربین[1] در ورودی مدل، می‌تواند وابستگی خروجی عمق به پارامترهای دوربین را کاهش داده و پایداری مقیاس فیزیکی را بهبود بخشد؟
- آیا بهبود عملکرد مدل در معیارهای متریک، بدون تخریب ساختار نسبی عمق و معیارهای غیرمتریک قابل دستیابی است؟

این سوالات، به‌طور مستقیم از محدودیت‌های مدل‌های تخمین عمق تک‌دوربینهٔ متداول ناشی می‌شوند که عموماً خروجی‌هایی با ابهام مقیاس[2] تولید می‌کنند و نسبت به تغییر پارامترهای دوربین حساس هستند.

[1] Camera Intrinsics
[2] Scale Ambiguity

۲.۲.۴ فرضیات تجربی

به‌منظور پاسخ‌گویی نظام‌مند به سوالات فوق، فرضیات تجربی زیر در این پژوهش تعریف و مورد آزمون قرار گرفته‌اند:

- **فرضیهٔ اول:** مدل پایهٔ Depth Anything V2 در معیارهای متریک، نسبت به نوع دیتاست و پارامترهای دوربین حساس بوده و تفاوت عملکرد معنادار آماری بین دیتاست‌های مختلف نشان می‌دهد.
- **فرضیهٔ دوم:** ادغام صریح ماتریس مشخصهٔ دوربین K در مدل پیشنهادی (دارای K)، منجر به کاهش معنادار خطا در معیارهای متریک و بهبود پایداری مقیاس فیزیکی خروجی عمق می‌شود.
- **فرضیهٔ سوم:** بهبود عملکرد متریک مدل پیشنهادی (دارای K)، بدون افت محسوس در معیارهای غیرمتریک و ساختار نسبی نقشهٔ عمق حاصل می‌شود.

۳.۲.۴ ارتباط فرضیات با بخش‌های فصل

برای شفافیت در روند ارزیابی، هر یک از فرضیات فوق در بخش مشخصی از این فصل مورد بررسی قرار می‌گیرد:

- فرضیهٔ اول در بخش ۴.۴ و به‌طور خاص در تحلیل عملکرد مدل پایهٔ Depth Anything V2 در دیتاست‌های مختلف بررسی می‌شود.
- فرضیهٔ دوم در بخش‌های مقایسهٔ کمی بین مدل پایهٔ Depth Anything V2 و مدل پیشنهادی (دارای K) و همچنین در تحلیل پایداری بین‌دیتاستی مدل پیشنهادی (دارای K) مورد ارزیابی قرار می‌گیرد.
- فرضیهٔ سوم در بخش تحلیل معیارهای غیرمتریک و نتایج کیفی، از طریق مقایسهٔ بصری نقشه‌های عمق بررسی خواهد شد.

بدین ترتیب، ساختار فصل به‌گونه‌ای طراحی شده است که پاسخ‌گویی به سوالات تحقیق و آزمون فرضیات به‌صورت مرحله‌به‌مرحله و مبتنی بر شواهد تجربی انجام گیرد.

۳.۴ تنظیمات آزمایش و پروتکل ارزیابی

در این بخش، تنظیمات آزمایش‌ها، دیتاست‌های مورد استفاده و پروتکل ارزیابی مدل‌ها تشریح می‌شود. هدف از این بخش، ایجاد یک پل منطقی میان چارچوب روش‌شناختی ارائه‌شده در فصل سوم و نتایج تجربی گزارش‌شده در ادامهٔ این فصل است، به‌گونه‌ای که تفسیر نتایج بر مبنای شرایط آزمایش‌ها به‌صورت شفاف و قابل بازتولید انجام گیرد.

۱.۳.۴ دیتاست‌های مورد استفاده

به‌منظور ارزیابی جامع عملکرد مدل‌ها در شرایط متنوع، از مجموعه‌ای از دیتاست‌های استاندارد و پرکاربرد در حوزهٔ تخمین عمق تک‌دوربینه استفاده شده است. این دیتاست‌ها از نظر نوع محیط، منبع داده و پارامترهای دوربین تنوع بالایی دارند و امکان بررسی پایداری مدل نسبت به تغییر شرایط تصویربرداری را فراهم می‌کنند.

- **VKITTI:** دیتاستی مصنوعی[1] شامل صحنه‌های بیرونی، تولیدشده بر پایهٔ دیتاست KITTI، با عمق مرجع دقیق و چگال.
- **NYUv2:** دیتاستی واقعی از محیط‌های داخلی[2]، ثبت‌شده با دوربین RGB، که عمق مرجع آن با استفاده از حسگرهای فعال به‌دست آمده است.
- **DIODE:** دیتاستی ترکیبی شامل صحنه‌های داخلی و خارجی، با تنوع بالا در نوع دوربین و شرایط تصویربرداری.
- **DrivingStereo:** دیتاستی واقعی از محیط‌های بیرونی[3] با عمق مرجع متراکم، استخراج‌شده از داده‌های استریو با کالیبراسیون دقیق دوربین.
- **CityScapes:** دیتاستی شهری و واقعی، که عمق مرجع آن از طریق تطبیق استریو و الگوریتم‌های بازسازی سه‌بعدی به‌دست آمده و دارای چگالی و کیفیت ناهمگن در نواحی مختلف تصویر است.

این تنوع دیتاست‌ها امکان بررسی رفتار مدل‌ها در شرایط مختلف تصویربرداری، تفاوت میدان دید، و تغییر پارامترهای هندسی دوربین را فراهم می‌سازد.

[1] Synthetic Dataset
[2] Indoor
[3] Outdoor

NYUv2 — داخلی، عمق از حسگر نوری، میدان دید $\approx 63^\circ$

Middlebury — داخلی، دو نمای کالیبره، وضوح بالا

Cityscapes — شهری، عمق از تطبیق استریو

KITTI — جاده‌ای، عمق پراکندهٔ لایدار

شکل ۱.۴: نمونه‌ای از تنوع شرایط تصویربرداری در مجموعه‌داده‌های به‌کاررفته. تفاوت این صحنه‌ها تنها در محتوا نیست: میدان دید، نسبت ابعاد، دامنهٔ عمق و سازوکار تولید عمق مرجع در آن‌ها اساساً متفاوت است (جدول ۳.۳). تصاویر با نسبت ابعاد اصلی خود نمایش داده شده‌اند تا همین تفاوت دیده شود؛ برای نمونه، تصویر KITTI با نسبت ابعاد حدود ۳٫۵ به ۱ و میدان دید افقی بازِ دوربین جاده‌ای، در برابر قاب ۴ به ۳ و صحنهٔ بستهٔ NYUv2 قرار می‌گیرد. همین ناهمگونی است که ارزیابی با یک مقیاس متریک ثابت را دشوار می‌کند و انگیزهٔ لحاظ صریح پارامترهای درونی دوربین را شکل می‌دهد. هر چهار تصویر مستقیماً از خودِ مجموعه‌داده‌ها برداشته شده‌اند (نه از مقالات)؛ KITTI با مجوز CC BY-NC-SA 3.0 و سه مجموعهٔ دیگر با مجوز استفادهٔ پژوهشی منتشر شده‌اند.

۲.۳.۴ نوع محیط‌ها و سناریوهای ارزیابی

دیتاست‌های مورد استفاده را می‌توان از منظر نوع محیط به سه دستهٔ اصلی تقسیم کرد:

- محیط‌های داخلی (NYUv2 و بخشی از DIODE)
- محیط‌های خارجی (DrivingStereo، CityScapes و بخشی از DIODE)
- داده‌های مصنوعی (VKITTI)

این تقسیم‌بندی امکان تحلیل پایداری مدل‌ها در سناریوهای مختلف، از فضاهای بسته با مقیاس محدود تا

صحنه‌های باز شهری با عمق‌های بزرگ، را فراهم می‌کند.

۳.۳.۴ معیارهای ارزیابی

برای ارزیابی عملکرد مدل‌ها، از ترکیبی از معیارهای غیرمتریک و متریک استفاده شده است تا هم ساختار نسبی عمق و هم مقیاس فیزیکی خروجی مورد بررسی قرار گیرد.

۱.۳.۳.۴ معیارهای غیرمتریک

معیارهای غیرمتریک[1] برای ارزیابی صحت نسبی عمق و ترتیب اجسام در صحنه به‌کار می‌روند و نسبت به مقیاس فیزیکی مطلق حساس نیستند. در این پژوهش، معیار AbsRel به‌عنوان نمایندهٔ این دسته استفاده شده است.

۲.۳.۳.۴ معیارهای متریک

معیارهای متریک[2] برای سنجش دقت عمق بر حسب واحدهای فیزیکی واقعی (مانند متر) به‌کار می‌روند و نسبت به خطاهای مقیاس حساس هستند. در این پژوهش، معیار SiLog به‌عنوان معیار اصلی ارزیابی عمق متریک استفاده شده و در برخی تحلیل‌ها، معیار RMSE نیز گزارش می‌شود.

۴.۳.۴ آزمون‌های آماری

به‌منظور بررسی معناداری تفاوت‌های مشاهده‌شده بین مدل‌ها و دیتاست‌ها، از آزمون‌های آماری استفاده شده است. در مقایسه‌های بین‌دیتاستی این فصل، آزمون t دو نمونه‌ای[3] برای مقایسهٔ میانگین خطاها به‌کار رفته و سطح معناداری برابر با ۰/۰۵ در نظر گرفته شده است.

در مطالعهٔ کنترل‌شدهٔ بخش ۵.۴ که در آن مدل‌های مختلف بر روی مجموعهٔ آزمون یکسانی ارزیابی می‌شوند، از آزمون‌های زوجی در سطح تک‌تصویر استفاده شده است: برای هر تصویر آزمون، خطای دو مدل به‌صورت زوجی محاسبه شده و معناداری تفاوت میانگین‌ها با آزمون t زوجی[4] سنجیده می‌شود. از آنجا که

[1] Non-metric Metrics
[2] Metric Metrics
[3] Two-sample t-test
[4] Paired t-test

توزیع خطاهای تک‌تصویری لزوماً نرمال نیست، آزمون ناپارامتری علامت‌دار ویلکاکسون[1] نیز به‌عنوان سنجهٔ استحکام در کنار آن گزارش می‌شود. طرح زوجی، واریانس ناشی از دشواری متفاوت تصاویر را حذف کرده و توان آماری آزمون را به‌طور قابل توجهی افزایش می‌دهد.

علاوه بر این، به‌منظور برآورد پایداری نتایج و کاهش وابستگی به فرضیات توزیعی، از بازنمونه‌گیری بوت‌استرپ[2] (با ده هزار بازنمونه بر روی تفاضل‌های زوجی خطا) برای محاسبهٔ بازه‌های اطمینان ۹۵٪ استفاده شده است. ترکیب این رویکردهای آماری امکان تحلیل دقیق‌تر تفاوت عملکرد مدل‌ها و اطمینان از اعتبار نتایج تجربی را فراهم می‌کند.

۴.۴ نتایج کمی: مقایسهٔ عددی مدل‌ها

در این بخش، نتایج کمی حاصل از ارزیابی مدل‌ها ارائه و تحلیل می‌شود. هدف اصلی، بررسی پایداری مدل پایهٔ Depth Anything V2 در دیتاست‌های مختلف و مقایسهٔ آن با مدل پیشنهادی (دارای K) از منظر معیارهای متریک و غیرمتریک است. تمامی نتایج گزارش‌شده در این بخش بر اساس اجرای یکسان پروتکل ارزیابی و با استفاده از آزمون‌های آماری ارائه‌شده در بخش ۳.۴ به‌دست آمده‌اند.

لازم به توضیح است که آزمایش‌های این بخش، مرحلهٔ نخست مطالعات تجربی این پژوهش را تشکیل می‌دهند و پیش از اصلاح تفکیک مجموعه‌های آموزش و آزمون NYUv2 (بخش ۳-۷) انجام شده‌اند؛ از این رو، مقادیر مطلق گزارش‌شده برای NYUv2 در این بخش باید صرفاً به‌عنوان مقایسهٔ نسبی میان مدل‌ها تفسیر شوند. مجموعهٔ نهایی و بدون نشت آزمایش‌ها، شامل مقایسهٔ کنترل‌شده با مدل پایهٔ Depth Anything V2 و مدل کنترل (فاقد K) هم‌داده و تحلیل آماری زوجی کامل، در بخش ۵.۴ ارائه می‌شود.

۱.۴.۴ عملکرد مدل پایهٔ Depth Anything V2 در دیتاست‌های مختلف

در گام نخست، عملکرد مدل پایهٔ Depth Anything V2 در دیتاست‌های مختلف مورد بررسی قرار می‌گیرد تا میزان پایداری آن نسبت به نوع دیتاست و پارامترهای دوربین ارزیابی شود. برای این منظور، معیار SiLog به‌عنوان معیار متریک اصلی استفاده شده است.

نتایج مقایسهٔ بین‌دیتاستی مدل پایهٔ Depth Anything V2 در جدول ۱.۴ ارائه شده است.

[1] Wilcoxon Signed-rank Test
[2] Bootstrap Resampling

جدول ۱.۴: مقایسهٔ معیار SiLog مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده در دیتاست‌های مختلف

مقایسهٔ دیتاست‌ها	اختلاف میانگین	p-value	نتیجه
VKITTI vs DrivingStereo	۰٫۰۳۹	$<$ ۰٫۰۵	معنادار
NYUv2 vs DrivingStereo	۰٫۱۰۱	$<$ ۰٫۰۵	معنادار
NYUv2 vs DIODE	۰٫۴۵۰	$<$ ۰٫۰۵	معنادار

همان‌طور که مشاهده می‌شود، مدل پایهٔ Depth Anything V2 در تمامی مقایسه‌های انجام‌شده تفاوت عملکرد معنادار آماری بین دیتاست‌ها نشان می‌دهد ($p <$ ۰٫۰۵). این رفتار بیانگر آن است که خروجی عمق مدل پایهٔ Depth Anything V2 به نوع دیتاست و در نتیجه به پارامترهای دوربین و شرایط تصویربرداری وابسته است. در نتیجه، می‌توان نتیجه گرفت که مدل پایهٔ Depth Anything V2 قادر به تولید عمق متریک پایدار در شرایط مختلف نیست.

این ناپایداری، حتی در مواردی که مدل از نظر ساختار نسبی عمق عملکرد مناسبی دارد، در معیارهای متریک به‌صورت اختلاف‌های قابل توجه در مقدار خطا ظاهر می‌شود.

۲.۴.۴ مقایسهٔ مدل پایهٔ Depth Anything V2 و مدل پیشنهادی (دارای K)

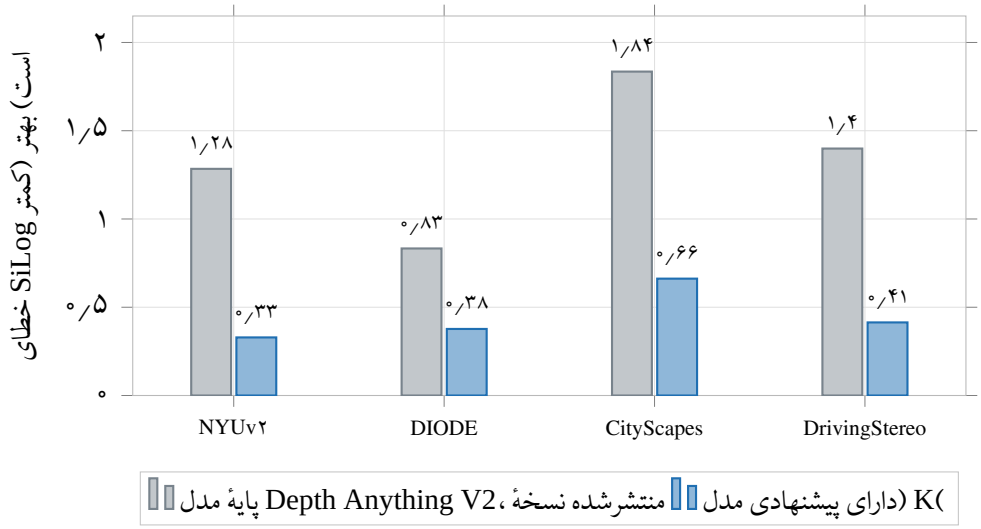
در ادامه، عملکرد مدل پایهٔ Depth Anything V2 با مدل پیشنهادی (دارای K) که در آن ماتریس مشخصهٔ دوربین K به‌صورت صریح در فرآیند تخمین عمق ادغام شده است، مقایسه می‌شود. تمرکز اصلی این مقایسه بر معیارهای متریک و به‌ویژه SiLog است.

جدول ۲.۴ نتایج مقایسهٔ مستقیم دو مدل در دیتاست‌های مختلف را نشان می‌دهد.

جدول ۲.۴: مقایسهٔ عملکرد مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده و مدل پیشنهادی (دارای K) بر اساس معیار SiLog

دیتاست	پایهٔ منتشرشده	پیشنهادی	بهبود (%)	p-value
NYUv2	۱٫۲۸۴	۰٫۳۲۹	۷۴٫۳۸	$\ll$ ۰٫۰۵
DIODE	۰٫۸۳۳	۰٫۳۷۷	۵۴٫۷۸	$\ll$ ۰٫۰۵
CityScapes	۱٫۸۳۵	۰٫۶۶۲	۶۳٫۹۰	$\ll$ ۰٫۰۵
DrivingStereo	۱٫۳۹۹	۰٫۴۱۴	۷۰٫۴۲	$\ll$ ۰٫۰۵

نتایج نشان می‌دهد که مدل پیشنهادی (دارای K) در تمامی دیتاست‌ها کاهش چشمگیر و معنادار آماری در مقدار خطای SiLog نسبت به مدل پایهٔ Depth Anything V2 دارد. درصد بهبود مشاهده‌شده بین حدود ۵۵ تا ۷۴ درصد متغیر بوده که بیانگر تأثیر قابل توجه ادغام اطلاعات هندسی دوربین بر تخمین عمق متریک است.



شکل ۲.۴: مقایسهٔ عملکرد متریک (معیار SiLog، کمتر بهتر است) میان مدل پایهٔ Depth Anything V2 و مدل پیشنهادی (دارای K) در چهار دیتاست با محیط و مشخصات دوربین متفاوت. مدل پیشنهادی (دارای K) در تمامی موارد خطای کمتری دارد و — مهم‌تر از مقدار مطلق بهبود — پراکندگی مقادیر آن میان دیتاست‌ها بسیار کمتر است، که همان پایداری بین‌دیتاستی مورد ادعای این پژوهش است. یادآوری می‌شود که مقادیر این نمودار به مرحلهٔ اکتشافی و پیش از اصلاح تفکیک آموزش/آزمون NYUv2 تعلق دارند و باید تنها به‌صورت مقایسهٔ نسبی تفسیر شوند؛ نتایج نهایی و بدون نشت در بخش ۵.۴ ارائه شده است.

این بهبود یکنواخت در دیتاست‌های داخلی، خارجی و مصنوعی نشان می‌دهد که مدل پیشنهادی (دارای K) قادر است وابستگی به توزیع داده و نوع دوربین را به‌طور مؤثری کاهش دهد.

۳.۴.۴ تحلیل پایداری بین‌دیتاستی مدل پیشنهادی (دارای K)

در این بخش، پایداری مدل پیشنهادی (دارای K) نسبت به تغییر دیتاست و دوربین بررسی می‌شود. برای این منظور، عملکرد مدل پیشنهادی (دارای K) بین جفت دیتاست‌های مختلف با استفاده از آزمون t و بازه‌های اطمینان بوت‌استرپ مقایسه شده است.

نتایج نشان می‌دهد که برخلاف مدل پایهٔ Depth Anything V2، تفاوت عملکرد مدل پیشنهادی (دارای K) در بسیاری از مقایسه‌های بین‌دیتاستی از نظر آماری معنادار نیست ($p > ۰٫۰۵$) و بازه‌های اطمینان محاسبه‌شده شامل مقدار صفر هستند.

این رفتار بیانگر آن است که مدل پیشنهادی (دارای K) توانسته است وابستگی خروجی عمق به پارامترهای دوربین را تا حد زیادی کاهش دهد و به عمق متریک پایدار دست یابد. بدین ترتیب، می‌توان نتیجه گرفت که ادغام صریح ماتریس مشخصهٔ دوربین نقش کلیدی در بهبود پایداری مقیاس در تخمین عمق تک‌دوربینه ایفا می‌کند و

پاسخ مثبت و مستقیمی به سوال اصلی پژوهش در خصوص امکان دستیابی به عمق متریک پایدار ارائه می‌دهد.

۵.۴ مطالعهٔ کنترل‌شده بر روی پروتکل استاندارد NYUv2

در مرحلهٔ دوم مطالعات تجربی این پژوهش، به‌منظور ارائهٔ مقایسه‌ای کاملاً منصفانه، بدون نشت داده و قابل انطباق با ادبیات موضوع، مجموعه‌ای از آزمایش‌های کنترل‌شده بر روی پروتکل استاندارد ارزیابی NYUv2 طراحی و اجرا شد. در این پروتکل که در اغلب پژوهش‌های مرجع از جمله Depth Anything V2 به‌کار می‌رود [۲، ۱۳]، ارزیابی بر روی زیرمجموعهٔ آزمون ۶۵۴ تصویری Eigen با نقشه‌های عمق خام حسگر به‌عنوان مرجع، برش استاندارد Eigen، و محدودهٔ عمق ۰٫۰۰۱ تا ۱۰ متر انجام می‌شود. برای مدل‌های نسبی، پیش از محاسبهٔ معیارها، تراز مقیاس و جابجایی به‌ازای هر تصویر در فضای عمق معکوس اعمال می‌گردد و سپس پیش‌بینی ترازشده به بازهٔ معتبر ارزیابی محدود می‌شود؛ برای مدل‌های متریک، خروجی مدل بدون هیچ‌گونه ترازی مستقیماً با مرجع مقایسه می‌شود. در بازتولید KITTI نیز مطابق جدول ۲ مقالهٔ مرجع از ۶۵۲ تصویر تقسیم Eigen و حالت بدون برش استفاده شد.

برای هر اندازهٔ انکودر، چهار مدل در شرایط یکسان مقایسه می‌شوند:

۱. مقادیر گزارش‌شده در مقالهٔ مرجع Depth Anything V2 (جدول ۲ مقاله)؛

۲. مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، ارزیابی‌شده با زنجیرهٔ ارزیابی این پژوهش؛

۳. مدل کنترل (فاقد K): معماری پایه بدون پارامترهای دوربین، فاین‌تیون‌شده بر روی همان دادهٔ آموزشی بدون نشت، با ابرپارامترها و برنامهٔ آموزشی کاملاً یکسان با مدل پیشنهادی (دارای K)؛

۴. مدل پیشنهادی (دارای K): همان تنظیمات، به‌علاوهٔ ورودی صریح ماتریس مشخصهٔ دوربین.

مقایسهٔ (۴) با (۲) میزان بهبود نسبت به مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده را می‌سنجد و مقایسهٔ (۴) با (۳) — که تنها در دسترسی به K تفاوت دارند — سهم خالص پارامترهای درونی دوربین را جدا می‌کند.

۱.۵.۴ ممیزی نقاط وارسی و مسیر رسیدن به آزمایش نهایی

هدف این بخش گزارش رتبه یا عملکرد نهایی مدل‌ها نیست. پرسش این بخش آن است که چرا پروتکل نهایی پژوهش به شکل فعلی طراحی شد؟ برای پاسخ به این پرسش، هر آزمایش به‌صورت یک زنجیرهٔ سه‌مرحله‌ای

گزارش می‌شود: نخست یک مشکل یا الگوی مشکوک مشاهده شد، سپس معنای آن بررسی شد و در پایان، بر پایهٔ آن یک تصمیم برای آزمایش بعدی گرفته شد. بنابراین، هر سطر جدول ۳.۴ باید از ستون «مشاهده» به ستون «برداشت» و سپس به ستون «اقدام بعدی» خوانده شود.

نخست، حدود بیست نقطهٔ وارسی (وزن‌های ذخیره‌شدهٔ مدل در حین یا پایان آموزش) ممیزی شدند. این نقاط وارسی اندازه‌های ViT-S و ViT-L، حالت‌های عمق نسبی و متریک، و ترکیب‌های مختلف داده را پوشش می‌دادند. در بیشتر اجراها، خطای AbsRel روی دادهٔ اعتبارسنجی درون‌دامنه‌ای بین ۰٫۰۵ و ۰٫۰۷ بود؛ در این معیار عدد کمتر بهتر است. بااین‌حال، دو اجرا در تمام دوره‌ها دقیقاً روی ۰٫۳۶۹۸ باقی مانده بودند و دو نقطهٔ وارسی نسبی ViT-L نیز برای همهٔ تصاویر خروجی صفر یا نامعتبر تولید می‌کردند. ثابت‌ماندن یک معیار در تمام دوره‌ها، همراه با خروجی نامعتبر، یک نتیجهٔ ضعیف معمولی نیست؛ نشانه‌ای از فروریزش آموزش یا ناسازگاری در ذخیره و بازیابی مدل است. ازاین‌رو، پایین‌بودن خطای اعتبارسنجی به‌تنهایی برای سالم‌بودن مدل یا توانایی تعمیم آن کافی در نظر گرفته نشد و هر ادعا با ارزیابی مستقل و بررسی مستقیم آماره‌های خروجی کنترل شد.

جدول ۳.۴: مسیر تصمیم‌گیری از نقاط وارسی اولیه تا آزمایش نهایی آگاه از دوربین. این جدول رتبه‌بندی مدل‌ها نیست؛ هر سطر نشان می‌دهد یک مشاهده چگونه به یک برداشت و سپس به تصمیم آزمایشی بعدی منجر شد.

گام	چه چیزی مشاهده شد؟	این مشاهده چه معنایی داشت؟	اقدام بعدی چه بود؟
۱. ممیزی اولیه	چند نقطهٔ وارسی خروجی صفر یا نامعتبر داشتند؛ بااین‌حال، منحنی اعتبارسنجی بعضی از آن‌ها عادی به نظر می‌رسید.	یک معیار اعتبارسنجی به‌تنهایی نمی‌تواند سلامت مدل را تضمین کند.	خط مبنای منتشرشده بازتولید و خروجی همهٔ مدل‌ها مستقل از فرایند آموزش بررسی شد.

ادامه در صفحهٔ بعد

ادامهٔ جدول ۳.۴

گام	چه چیزی مشاهده شد؟	این مشاهده چه معنایی داشت؟	اقدام بعدی چه بود؟
۲. ارزیابی صفرنمونه	پس از فاین‌تیون روی چند مجموعه‌دادهٔ محدود، خطای KITTI تقریباً از ۰٫۰۸۳ به ۰٫۲۱۴ افزایش یافت؛ این افت در مدل‌های پیشنهادی و کنترل تقریباً یکسان بود.	فاین‌تیون محدود، توان تعمیم مدل بنیادین را تضعیف کرده بود و افزودن K علت این افت نبود.	مسیر ارزیابی صفرنمونه کنار گذاشته شد و مطالعه به پروتکل استاندارد درون‌دامنه‌ای منتقل شد.
۳. پاک‌سازی داده	۵۲۷ تصویر متعلق به بخش آزمون NYUv2 در دادهٔ آموزش اولیه پیدا شد.	نتیجهٔ آن آموزش به نشت داده آلوده بود و نمی‌توانست شاهد معتبری برای تعمیم باشد.	تمام تصاویر تقسیم آزمون Eigen پیش از ساخت مجموعه‌های آموزش و اعتبارسنجی حذف شدند.
۴. آزمایش نسبی ViT-S	پس از حذف نشت، مدل پیشنهادی از پایهٔ منتشرشده و کنترل بهتر شد (نسخهٔ نسبی، ViT-S).	نخستین شاهد معتبر به دست آمد؛ اما هنوز معلوم نبود نتیجه در مدل بزرگ‌تر نیز تکرار می‌شود.	همان آزمایش برای ViT-L تکرار شد.
۵. فروریزش ViT-L	مدل بزرگ‌تر خروجی تقریباً صفر تولید کرد، درحالی‌که مقدار تابع هزینه ظاهراً غیرعادی نبود.	نمایش هدفِ آموزش با خروجی مورد انتظار ناسازگار بود؛ بنابراین، منحنی هزینه مشکل را آشکار نمی‌کرد.	هدف از عمق مستقیم به عمق معکوس تغییر کرد و نوع فضای هدف نیز در نقطهٔ وارسی ذخیره شد.

ادامه در صفحهٔ بعد

ادامهٔ جدول ۳.۴

گام	چه چیزی مشاهده شد؟	این مشاهده چه معنایی داشت؟	اقدام بعدی چه بود؟
۶. زوم‌ـبرش هندسی	در برش معمولی تصویر، میدان دید تغییر می‌کرد ولی K ثابت می‌ماند و درنتیجه از نظر هندسی نادرست می‌شد. پس از اصلاح K، هر دو مدل بهتر شدند، اما مدل پیشنهادی در ارزیابی نسبی برتری روشنی نداشت.	تنوع هندسی برای هر دو مدل مفید بود؛ ولی تراز نسبی می‌توانست اثر اطلاعات مقیاس موجود در K را حذف کند.	ارزیابی به حالت متریک و بدون تراز منتقل شد.
۷. ارزیابی متریک	با حذف تراز نسبی، هرچه تنوع میدان دید در آموزش بیشتر شد، استفادهٔ مدل از اطلاعات دوربین نیز منظم‌تر افزایش یافت.	اثر K عمدتاً به بازیابی مقیاس متریک مربوط بود و پروتکل نسبی آن را پنهان می‌کرد.	آزمایشی طراحی شد که مقیاس بدون دانستن کانون دوربین قابل بازیابی نباشد.
۸. لرزش کانونی	در طراحی اولیه، شیب پاسخ مقیاس به تغییر K فقط ۰٫۰۶ بود؛ پس از افزودن لرزش کانونی به دادهٔ آموزش، شیب به ۰٫۹۳-۰٫۹۷ رسید. مقدار ایده‌آل این شیب ۱ است.	شیب ۰٫۰۶ یعنی مدل تقریباً K را نادیده می‌گرفت؛ شیب نزدیک ۱ یعنی مقیاس خروجی تقریباً متناسب با کانون ورودی تغییر می‌کرد.	لرزش کانونی در طراحی نهایی حفظ شد و ضرورت استفاده از K با آزمایش‌های علّی بررسی شد.

ادامه در صفحهٔ بعد

ادامهٔ جدول ۳.۴

گام	چه چیزی مشاهده شد؟	این مشاهده چه معنایی داشت؟	اقدام بعدی چه بود؟
۹. فروکاست و حذف K	در آزمایش قبلی، هم لرزش کانونی و هم وزن زیان λ تغییر کرده بودند؛ بنابراین، عامل اصلی بهبود هنوز قطعی نبود.	هم‌بستگی میان K و بهبود عملکرد برای ادعای رابطهٔ علّی کافی نبود.	با ثابت نگه‌داشتن سایر عوامل نشان داده شد لرزش کانونی عامل بهبود است و جایگزینی یا حذف K صحیح، مقیاس پیش‌بینی را مختل می‌کند.

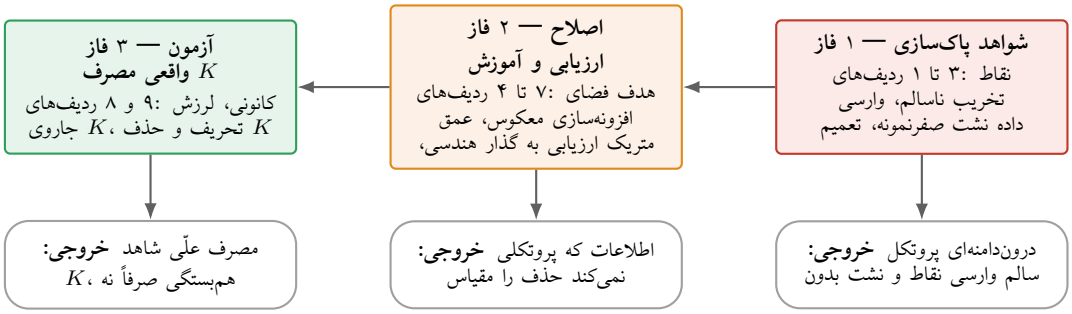
جدول را می‌توان در سه فاز خلاصه کرد. در **فاز نخست، پاک‌سازی شواهد** (ردیف‌های ۱ تا ۳)، نقاط وارسی ناسالم، تعمیم صفرنمونهٔ تخریب‌شده و نشت داده شناسایی شدند. خروجی این فاز یک پروتکل درون‌دامنه‌ای بدون نشت بود که در آن نتیجهٔ مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) قابل مقایسه است.

در **فاز دوم، اصلاح آموزش و ارزیابی** (ردیف‌های ۴ تا ۷)، ابتدا امکان کسب بهبود بدون نشت تأیید شد، سپس فروریزش مدل بزرگ‌تر با تغییر هدف به عمق معکوس رفع گردید. همچنین روشن شد که ارزیابی نسبی، با ترازکردن جداگانهٔ هر تصویر، اطلاعات مقیاس را حذف می‌کند. به همین دلیل، اثر پارامترهای دوربین باید در حالت متریک و بدون تراز سنجیده شود.

در **فاز سوم، آزمون استفاده‌واقعی از دوربین** (ردیف‌های ۸ و ۹)، لرزش کانونی باعث شد مقیاس صحنه از ظاهر تصویر به‌تنهایی قابل حدس نباشد. در این شرایط، مدل برای بازیابی مقیاس ناچار بود از K استفاده کند. افزایش شیب جاروی K از ۰٫۰۶ به بازهٔ ۰٫۹۳ تا ۰٫۹۷ نشان داد پاسخ مدل از حالت تقریباً بی‌تفاوت به حالتی نزدیک به پاسخ هندسی ایده‌آل رسیده است. در نهایت، آزمایش‌های فروکاست و حذف K نشان دادند که این اثر صرفاً هم‌بستگی نیست: وقتی K صحیح حذف یا با مقدار نادرست جایگزین شد، مقیاس خروجی نیز به‌طور قابل پیش‌بینی مختل شد.

شکل ۳.۴ این سه فاز و شرط خروجی هر یک را نشان می‌دهد.

بنابراین، پیام اصلی جدول آن نیست که هر ردیف یک مدل بهتر معرفی می‌کند. پیام جدول این است که اعتبار نتیجهٔ نهایی به چهار شرط وابسته است: دادهٔ بدون نشت، نقطهٔ وارسی سالم، ارزیابی متریکِ بدون تراز، و آزمایشی که در آن K برای بازیابی مقیاس واقعاً ضروری باشد.



شکل ۳.۴: سه فاز مسیر تجربی پژوهش و شرط خروجی هر فاز. اعتبار نتیجهٔ نهایی به هر چهار شرط وابسته است: دادهٔ بدون نشت، نقطهٔ وارسی سالم، ارزیابی متریکِ بدون تراز، و آزمایشی که در آن K برای بازیابی مقیاس واقعاً ضروری باشد. جزئیات هر گام در جدول ۳.۴ آمده است.

۲.۵.۴ راستی‌آزمایی زنجیرهٔ‌ارزیابی: بازتولید نتایج مدل پایهٔ Depth Anything V2

پیش از هرگونه مقایسه، صحت زنجیرهٔ ارزیابی از طریق بازتولید نتایج گزارش‌شدهٔ مقالهٔ مرجع راستی‌آزمایی شد. جدول ۴.۴ نشان می‌دهد که نتایج به‌دست‌آمده با زنجیرهٔ ارزیابی این پژوهش، با دقت مطلوبی بر مقادیر مقاله منطبق است؛ اختلاف‌های جزئی در محدودهٔ تلورانس متعارف بازتولید این محک قرار دارند.

جدول ۴.۴: راستی‌آزمایی زنجیرهٔ ارزیابی: بازتولید نتایج مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ نسبی، بر روی پروتکل استاندارد. هر معیار در ستونی مستقل گزارش شده است.

محک	انکودر	مقالهٔ مرجع		ارزیابی این پژوهش	
		↓AbsRel	$\delta_1 \uparrow$	↓AbsRel	$\delta_1 \uparrow$
NYUv2	ViT-S	۰٫۰۵۳	۰٫۹۷۳	۰٫۰۵۱	۰٫۹۷۳
NYUv2	ViT-L	۰٫۰۴۵	۰٫۹۷۹	۰٫۰۴۳	۰٫۹۷۹
KITTI	ViT-S	۰٫۰۷۸	۰٫۹۳۶	۰٫۰۸۳	۰٫۹۳۴
KITTI	ViT-L	۰٫۰۷۴	۰٫۹۴۶	۰٫۰۷۶	۰٫۹۴۸

۳.۵.۴ ناکارآمدی فاین‌تیون برای ارزیابی برون‌دامنه‌ای

نخستین یافتهٔ این مرحله، نتیجه‌ای منفی اما تعیین‌کننده در طراحی ادامهٔ آزمایش‌ها بود: فاین‌تیون مدل پایهٔ Depth Anything V2 بر روی ترکیب مجموعه‌داده‌های محدود (نظیر VKITTI2، Cityscapes، DIODE و NYUv2)، توانایی تعمیم صفرنمونهٔ مدل را به‌شدت تخریب می‌کند. نتایج دقیق در جدول ۵.۴ آمده است.

افزایش KITTI AbsRel از حدود ۰٫۰۸۳ به ۰٫۲۱۴ هم در مدل پیشنهادی (دارای K) و هم در مدل کنترل

جدول ۵.۴: ارزیابی بدون نشت نسخه‌های فاین‌تیون‌شدهٔ چندمجموعه‌ای در حالت صفرنمونه، در مقایسه با مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ نسبی ViT-S. هر معیار در ستونی مستقل گزارش شده است.

محک	مدل	دادهٔ فاین‌تیون	↓AbsRel	δ_1 ↑
KITTI	پیشنهادی	چندمجموعه‌ای	۰٫۲۱۳۹	۰٫۶۱۸۹
KITTI	پایهٔ منتشرشده	—	۰٫۰۸۲۹	۰٫۹۳۴۱
KITTI	کنترل	چندمجموعه‌ای	۰٫۲۱۳۲	۰٫۶۱۷۹
KITTI	پایهٔ منتشرشده	—	۰٫۰۸۲۹	۰٫۹۳۴۱
NYUv2	پیشنهادی	چندمجموعه‌ای، بدون NYUv2	۰٫۰۶۵۷	۰٫۹۵۷۸
NYUv2	پایهٔ منتشرشده	—	۰٫۰۵۰۸	۰٫۹۷۳۳

(فاقد K) رخ می‌دهد؛ اختلاف ۰٫۲۱۳۹ و ۰٫۲۱۳۲ در حد نوفه است. بنابراین، پارامترهای دوربین نه علت این افت‌اند و نه راه‌حل آن. این مشاهده نشان می‌دهد که مدل بنیادین آموزش‌دیده بر ده‌ها میلیون تصویر، با فاین‌تیون بر چند هزار تصویر کوچک یا مصنوعی از بازنمایی عمومی خود فاصله می‌گیرد. بر همین اساس، مطابق رویهٔ مقالهٔ مرجع، ادامهٔ مطالعه به فاین‌تیون بر بخش آموزشی یک محک و ارزیابی بر بخش آزمون همان محک منتقل شد.

۴.۵.۴ نتایج پروتکل نسبی: مقایسهٔ چهارگانه

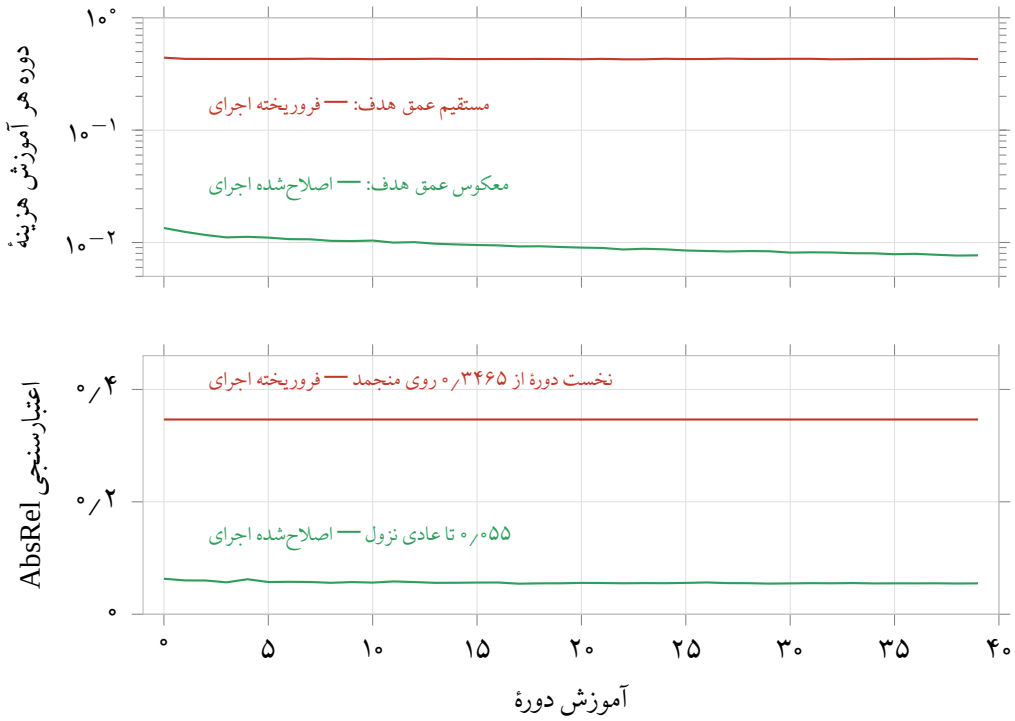
جدول‌های ۶.۴ و ۷.۴ نتایج مقایسهٔ چهارگانه را برای دو اندازهٔ انکودر نشان می‌دهند. هر دو مدل فاین‌تیون‌شده بر روی تفکیک بدون نشت (۶۳۶ تصویر آموزش) و با دستورالعمل یکسان (۴۰ دوره، نرخ یادگیری ۵×۱۰^{-۶}، انکودر فریزشده، هدف عمق معکوس مطابق بخش ۳-۷) آموزش دیده‌اند.

جدول ۶.۴: مقایسهٔ چهارگانه بر روی مجموعهٔ آزمون استاندارد NYUv2 (۶۵۴ تصویر، مرجع عمق خام)، انکودر ViT-S، پروتکل نسبی با تراز مقیاس و جابجایی.

مدل	دادهٔ فاین‌تیون	↓AbsRel	δ_1 ↑	↓RMSE	↓SqRel	$\log_{10}$ ↓
مقالهٔ مرجع (جدول ۲)	—	۰٫۰۵۳	۰٫۹۷۳	—	—	—
پایهٔ منتشرشده	—	۰٫۰۵۰۸	۰٫۹۷۳۳	۰٫۲۴۱۴	۰٫۰۲۴۵	۰٫۰۲۲۲
کنترل	NYUv2	۰٫۰۵۰۷	۰٫۹۷۴۱	۰٫۲۰۸۸	۰٫۰۱۸۶	۰٫۰۲۱۸
پیشنهادی	NYUv2	۰٫۰۵۰۱	۰٫۹۷۴۵	۰٫۲۰۶۴	۰٫۰۱۸۲	۰٫۰۲۱۶

پیش از دستیابی به نتیجهٔ ViT-L جدول بعد، دو اجرای مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) با همان دستورالعمل اولیه فروریختند: AbsRel اعتبارسنجی از دورهٔ نخست روی حدود ۰٫۳۴۶ ثابت ماند

و خروجی شبکه در تمام پیکسل‌ها دقیقاً صفر شد. کاهش ضریب نرخ یادگیری سر شبکه از ۱۰ به ۱ و افزودن گرم‌سازی نیز مشکل را رفع نکرد. بازتولید ابزارگذاری‌شده نشان داد سهم پیکسل‌های مرده در تنها ۱۸ گام از حدود ۰٫۰۰۱ به ۱ رسید. پس از انتقال هدف به عمق معکوس، یک اجرای آزمایشی ۱۵۰ گامی هیچ پیکسل مرده‌ای نداشت و هزینهٔ آن ۲۰ تا ۵۰ برابر کمتر بود؛ سپس آموزش کامل تکرار شد. بنابراین نتیجهٔ ViT-L زیر مربوط به نسخهٔ اصلاح‌شدهٔ فضای هدف است، نه اجرای فروریخته. امضای این پدیده — هزینهٔ آموزش ظاهراً سالم در کنار معیار اعتبارسنجی منجمد — در شکل ۴.۴ آمده است.



شکل ۴.۴: امضای فروریزش خاموش توضیح‌داده‌شده در بخش ۶.۷.۳، در دو اجرای کامل ۴۰ دوره‌ای ViT-L که تنها در فضای هدف تفاوت دارند. **بالا:** هزینهٔ آموزش اجرای فروریخته (قرمز) هموار و نزولی است — از ۰٫۴۴۲ به ۰٫۴۲۹ — و اگر تنها همین منحنی پایش شود، آموزش کاملاً سالم به نظر می‌رسد؛ زیرا با خروجی ثابت، مقدار هزینه صرفاً تابع آماره‌های هر دسته است. **پایین:** معیار اعتبارسنجی همان اجرا از دورهٔ نخست روی ۰٫۳۴۶۵ منجمد می‌ماند و تا پایان آموزش حتی در رقم چهارم اعشار تغییر نمی‌کند. نکتهٔ گویا آنکه در این حالت، مقادیر اعتبارسنجی دو بازوی مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) بیت‌به‌بیت یکسان است، چون خروجی هر دو شبکه ثابت صفر شده و دیگر به ورودی — و در نتیجه به K — وابسته نیست. پس از انتقال هدف به فضای عمق معکوس (سبز)، هر دو منحنی رفتار عادی می‌یابند. این شکل نشان می‌دهد که چرا پایش هزینهٔ آموزش به‌تنهایی برای تشخیص سلامت آموزش کافی نیست.

برای تکمیل گزارش اندازهٔ اثر، جدول ۸.۴ تفاضل زوجی و بازهٔ اطمینان بوت‌استرپ چهار مقایسهٔ اصلی ViT-S را ارائه می‌کند. علامت مثبت به‌معنای خطای بیشتر مدل مرجع یا مدل کنترل (فاقد K) نسبت به مدل

جدول ۷.۴: مقایسهٔ چهارگانه بر روی مجموعهٔ آزمون استاندارد NYUv2، انکودر ViT-L، پروتکل نسبی.

مدل	دادهٔ فاین‌تیون	↓AbsRel	δ_1 ↑	↓RMSE
مقالهٔ مرجع (جدول ۲)	—	۰٫۰۴۵	۰٫۹۷۹	—
پایهٔ منتشرشده	—	۰٫۰۴۲۵	۰٫۹۷۸۹	۰٫۲۱۱۹
کنترل	NYUv2	۰٫۰۳۷۸	۰٫۹۸۶۲	۰٫۱۹۲۲
پیشنهادی	NYUv2	۰٫۰۳۷۷	۰٫۹۸۶۱	۰٫۱۹۱۳

پیشنهادی (دارای K) است.

جدول ۸.۴: اندازهٔ اثر زوجی مدل پیشنهادی (دارای K) ViT-S بر ۶۵۴ تصویر NYUv2.

مقایسه	معیار	میانگین تفاضل	بازهٔ اطمینان ۹۵٪
پایهٔ منتشرشده	RMSE	+۰٫۰۳۵۰	[+۰٫۰۲۷۰, +۰٫۰۴۳۸]
پایهٔ منتشرشده	AbsRel	+۰٫۰۰۰۷	[−۰٫۰۰۰۸, +۰٫۰۰۲۲]
کنترل	RMSE	+۰٫۰۰۲۴	[+۰٫۰۰۱۳, +۰٫۰۰۳۴]
کنترل	AbsRel	+۰٫۰۰۰۶	[+۰٫۰۰۰۲, +۰٫۰۰۱۰]

برای سنجش معناداری آماری تفاوت‌ها، از آزمون‌های زوجی در سطح تک‌تصویر (آزمون t زوجی و آزمون علامت‌دار ویلکاکسون بر روی $n = ۶۵۴$ زوج) استفاده شده است؛ نتایج در جدول ۹.۴ آمده است.

جدول ۹.۴: آزمون‌های زوجی معناداری بر روی ۶۵۴ تصویر آزمون NYUv2 (پروتکل نسبی)؛ مقایسهٔ مدل پیشنهادی (دارای K) با مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده و مدل کنترل (فاقد K).

انکودر	مدل مرجع	معیار	p (t زوجی)	p (ویلکاکسون)
ViT-S	پایهٔ منتشرشده	AbsRel	۰٫۴۰	۰٫۰۹
ViT-S	پایهٔ منتشرشده	RMSE	$۲٫۸ \times ۱۰^{-۱۵}$	۵×۱۰^{-۱۸}
ViT-S	پایهٔ منتشرشده	δ_1	۰٫۳۴	۰٫۰۵
ViT-S	کنترل	AbsRel	$۱٫۳ \times ۱۰^{-۳}$	$۵٫۰ \times ۱۰^{-۳}$
ViT-S	کنترل	RMSE	$۱٫۱ \times ۱۰^{-۵}$	$۳٫۵ \times ۱۰^{-۵}$
ViT-S	کنترل	δ_1	۰٫۲۹	۰٫۶۰
ViT-L	پایهٔ منتشرشده	AbsRel	$۲٫۵ \times ۱۰^{-۸}$	$۸٫۹ \times ۱۰^{-۶}$
ViT-L	پایهٔ منتشرشده	RMSE	$۱٫۰ \times ۱۰^{-۷}$	$۱٫۱ \times ۱۰^{-۹}$
ViT-L	پایهٔ منتشرشده	δ_1	$۴٫۱ \times ۱۰^{-۶}$	$۱٫۲ \times ۱۰^{-۱۲}$
ViT-L	کنترل	AbsRel	۰٫۷۷	۰٫۱۸
ViT-L	کنترل	RMSE	۰٫۲۹	۰٫۵۰
ViT-L	کنترل	δ_1	۰٫۸۱	۰٫۰۴

سه نتیجهٔ اصلی از این جدول‌ها قابل استخراج است. نخست، مدل پیشنهادی (دارای K) ViT-L در هر سه معیار، به‌صورت معنادار آماری از مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده بهتر است (کاهش AbsRel از ۰٫۰۴۲۵ به ۰٫۰۳۷۷ و کاهش ۹٫۷ درصدی RMSE، همگی با $p < ۱۰^{-۵}$). دوم، در اندازهٔ

ViT-S، سهم خالص پارامترهای دوربین در برابر مدل کنترل (فاقد K) هم‌داده، کوچک اما از نظر آماری معنادار است AbsRel) با $p = ۱/۳ \times ۱۰^{-۳}$ و RMSE با $p = ۱/۱ \times ۱۰^{-۵}$). سوم — و از منظر علمی آموزنده — در اندازهٔ ViT-L تفاوت مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) در پروتکل نسبی از نظر آماری معنادار نیست؛ تحلیل این پدیده در ادامهٔ همین بخش و در بخش بحث ارائه می‌شود.

اندازهٔ اثر ViT-L در برابر مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده نیز روشن است. برای AbsRel، میانگین تفاضل زوجی و بازهٔ اطمینان ۹۵٪ به‌ترتیب ۰/۰۰۴۸+ و [۰/۰۰۶۶+, ۰/۰۰۳۲+] است. برای RMSE، مقادیر متناظر برابر ۰/۰۲۰۵+ و [۰/۰۲۸۲+, ۰/۰۱۳۳+] هستند. در مقابل، آزمون‌های مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) برای AbsRel و RMSE به‌ترتیب $p = ۰/۱۸$ و $p = ۰/۵۰$ در ویلکاکسون هستند و تفاوت معناداری میان دو بازو در این پروتکل نشان نمی‌دهند.

۵.۵.۴ اثر افزونه‌سازی هندسی و یک نتیجهٔ منفی آموزنده

فرضیهٔ محتمل برای عدم مشاهدهٔ سهم پارامترهای دوربین در ViT-L، ثابت‌بودن بردار $\mathbf{k}$ در مجموعه‌دادهٔ تک‌دوربینهٔ NYUv2 بود. برای آزمون این فرضیه، زوج مدل ViT-L با افزونه‌سازی هندسی آگاه از دوربین (بخش ۳-۷، بازهٔ میدان دید ۴۱ تا ۶۳ درجه) بازآموزی شد. نتایج (جدول ۱۰.۴) دو یافته به همراه داشت: نخست، خودِ افزونه‌سازی هندسی برای هر دو مدل سودمندی چشمگیری دارد (کاهش AbsRel مدل کنترل (فاقد K) از ۰/۰۳۷۸ به ۰/۰۳۵۹ با $p = ۱/۲ \times ۱۰^{-۱۳}$ و مدل پیشنهادی (دارای K) از ۰/۰۳۷۷ به ۰/۰۳۶۲ با $p = ۱/۲ \times ۱۰^{-۸}$) و بهترین نتایج این پژوهش در پروتکل نسبی را رقم می‌زند. دوم، حتی با وجود سیگنال هندسی متنوع و صحیح، مدل پیشنهادی (دارای K) در پروتکل نسبی بر مدل کنترل (فاقد K) برتری نیافت (و در AbsRel با اختلاف بسیار جزئی ۰/۰۰۰۳ ولی معنادار، $p = ۸/۱ \times ۱۰^{-۳}$، پشت سر آن قرار گرفت).

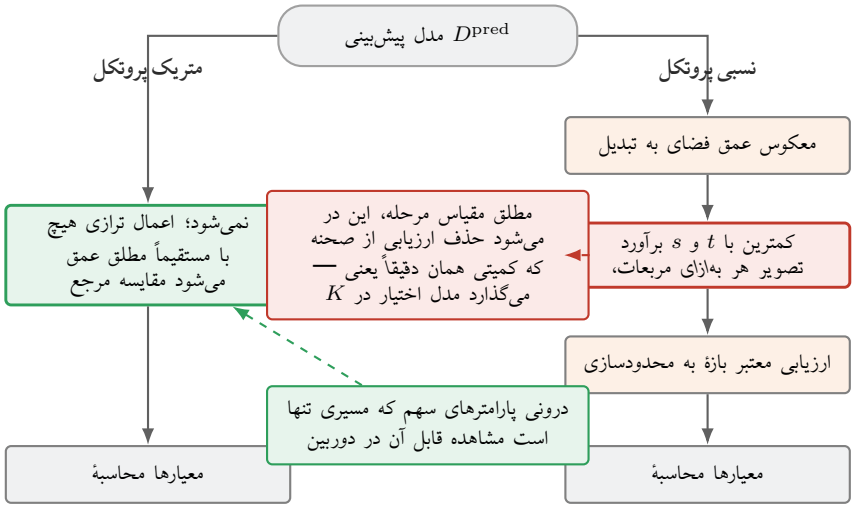
جدول ۱۰.۴: اثر افزونه‌سازی هندسی آگاه از دوربین بر زوج مدل ViT-L در پروتکل نسبی NYUv2.

مدل	↓AbsRel	$\delta_۱$ ↑	↓RMSE
کنترل	۰/۰۳۵۹	۰/۹۸۷۲	۰/۱۸۴۹
پیشنهادی	۰/۰۳۶۲	۰/۹۸۶۹	۰/۱۸۸۱

این نتیجهٔ منفی، تبیینی اصولی دارد که یکی از یافته‌های مفهومی این پژوهش محسوب می‌شود: در پروتکل ارزیابی نسبی، پیش از محاسبهٔ هر معیار، یک تراز مقیاس و جابجایی به‌ازای هر تصویر میان پیش‌بینی و مرجع اعمال می‌شود. اطلاعات اصلی که میدان دید و فاصلهٔ کانونی در اختیار مدل قرار می‌دهند، دقیقاً همان مقیاس مطلق صحنه است؛ کمیتی که سازوکار تراز، آن را از ارزیابی حذف می‌کند. به بیان دیگر، پروتکل نسبی به‌صورت

ساختاری نسبت به سهم اصلی پارامترهای درونی دوربین «کور» است. این تحلیل پیش‌بینی می‌کند که سهم پارامترهای دوربین باید در ارزیابی متریک (بدون تراز) آشکار شود؛ پیش‌بینی‌ای که در زیربخش بعد به‌صورت تجربی آزموده می‌شود.

شکل ۵.۴ این استدلال را به‌صورت تصویری نشان می‌دهد: دو پروتکل ارزیابی، تنها در یک مرحله تفاوت دارند و دقیقاً همان مرحله است که اطلاعات مورد بحث را از ارزیابی حذف می‌کند.



شکل ۵.۴: دو پروتکل ارزیابی به‌کاررفته در این پژوهش. تفاوت آن‌ها تنها در یک مرحله است: تراز مقیاس و جابجایی به‌ازای هر تصویر. اما همان یک مرحله، مؤلفهٔ مقیاس مطلق را از ارزیابی حذف می‌کند و مقیاس مطلق دقیقاً همان اطلاعاتی است که فاصلهٔ کانونی و میدان دید در اختیار مدل قرار می‌دهند. به این معنا، پروتکل نسبی به‌صورت ساختاری نسبت به سهم اصلی پارامترهای درونی دوربین نابیناست؛ و این توضیح می‌دهد که چرا نتیجهٔ جدول ۱۰.۴ یک شاهد علیه فرضیهٔ پژوهش نیست، بلکه پیامد قابل پیش‌بینی سازوکار ارزیابی است.

۶.۵.۴ ارزیابی متریک: آشکارشدن سهم پارامترهای درونی دوربین

در این مرحله، زوج مدل ViT-L در حالت متریک (تابع هزینهٔ SiLog بر عمق مطلق، مطابق فصل سوم) و با افزونه‌سازی هندسی آموزش دیده و بر روی مجموعهٔ آزمون استاندارد، بدون هیچ‌گونه تراز مقیاس، ارزیابی شد. مرجع مقایسه در این حالت، مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ متریک (فاین‌تیون‌شده بر Hypersim برای صحنه‌های داخلی) است. آزمایش در دو دور با دو شدت افزونه‌سازی انجام شد: دور نخست با بازهٔ میدان دید ۴۱ تا ۶۳ درجه و دور دوم با بازهٔ وسیع‌تر ۳۳ تا ۶۳ درجه.

نتایج جدول ۱۱.۴ و آزمون‌های زوجی متناظر (جدول ۱۲.۴) سه یافتهٔ کلیدی دارند:

نخست، هر دو مدل فاین‌تیون‌شده، مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ متریک را با

جدول ۱۱.۴: ارزیابی متریک (عمق مطلق، بدون تراز) بر روی مجموعهٔ آزمون استاندارد NYUv2، انکودر ViT-L. دور ۱: افزونه‌سازی با میدان دید ۴۱-۶۳ درجه؛ دور ۲: ۳۳-۶۳ درجه.

مدل	دادهٔ آموزش	دور	↓AbsRel	δ_1 ↑	↓RMSE
پایهٔ منتشرشده	Hypersim	—	۰٫۲۰۷۳	۰٫۶۹۱۹	۰٫۶۰۶۳
کنترل	NYUv2	۱	۰٫۰۷۹۶	۰٫۹۵۸۳	۰٫۲۹۱۹
پیشنهادی	NYUv2	۱	۰٫۰۷۹۱	۰٫۹۵۹۳	۰٫۲۸۸۲
کنترل	NYUv2	۲	۰٫۰۸۲۶	۰٫۹۵۴۶	۰٫۲۹۹۸
پیشنهادی	NYUv2	۲	۰٫۰۸۱۷	۰٫۹۵۶۹	۰٫۲۹۴۴

جدول ۱۲.۴: آزمون‌های زوجی معناداری در ارزیابی متریک ($n = ۶۵۴$)؛ مقایسهٔ مدل پیشنهادی (دارای K) با مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، نسخهٔ متریک، و مدل کنترل (فاقد K).

دور	مدل مرجع	معیار	p (t زوجی)	p (ویلکاکسون)
۱	پایهٔ منتشرشده	AbsRel	$۴٫۸ \times ۱۰^{-۱۳۳}$	$۸٫۴ \times ۱۰^{-۹۶}$
۱	پایهٔ منتشرشده	RMSE	$۶٫۷ \times ۱۰^{-۱۰۰}$	$۵٫۶ \times ۱۰^{-۸۷}$
۱	کنترل	AbsRel	۰٫۲۶	۰٫۵۳
۱	کنترل	RMSE	$۴٫۹ \times ۱۰^{-۳}$	$۹٫۳ \times ۱۰^{-۳}$
۲	کنترل	AbsRel	۰٫۰۸۹	۰٫۰۴۰
۲	کنترل	RMSE	$۶٫۵ \times ۱۰^{-۵}$	$۱٫۲ \times ۱۰^{-۴}$

فاصلهٔ بسیار زیاد پشت سر می‌گذارند (کاهش AbsRel از ۰٫۲۰۷ به حدود ۰٫۰۸، معادل بهبود ۲٫۶ برابری، با p در مرتبهٔ $۱۰^{-۱۳۳}$)؛ این نتیجه هم‌راستا با گزارش خود مقالهٔ مرجع است که مقیاس متریک مدل Hypersim به صحنه‌های واقعی به‌خوبی منتقل نمی‌شود. دوم، برخلاف پروتکل نسبی، در ارزیابی متریک مدل پیشنهادی (دارای K) در هر دو دور و در تمامی معیارها از مدل کنترل (فاقد K) هم‌داده پیشی می‌گیرد و این برتری در معیار RMSE در هر دو دور معنادار آماری است ($p = ۴٫۹ \times ۱۰^{-۳}$ در دور ۱ و $p = ۶٫۵ \times ۱۰^{-۵}$ در دور ۲). سوم — و مهم‌تر — با افزایش شدت تنوع میدان دید از دور ۱ به دور ۲، فاصلهٔ مدل پیشنهادی (دارای K) از مدل کنترل (فاقد K) در همهٔ معیارها افزایش می‌یابد و معناداری AbsRel نیز به آستانهٔ تصمیم می‌رسد (ویلکاکسون $p = ۰٫۰۴۰$). این الگوی پاسخ به دوز[1] — یعنی رشد سهم پارامترهای دوربین متناسب با میزان ابهام هندسی موجود در داده — که در دو اجرای مستقل آموزش تکرار شده است، شاهد تجربی مهمی بر سودمندی واقعی ورودی پارامترهای درونی دوربین است. زیربخش بعد نشان می‌دهد که با رساندن ابهام هندسی به حد نهایی خود، این سهم از یک برتری آماریِ مرزی به یک اثر قاطع و همه‌جانبه تبدیل می‌شود.

[1] Dose–Response Pattern

۷.۵.۴ دور سوم — افزونه‌سازی لرزش کانونی: جهش سهم پارامترهای دوربین

نقطهٔ اوج الگوی پاسخ به دوز، به‌کارگیری افزونه‌سازی لرزش کانونی (بخش ۸.۷.۳) است که مقیاس مطلق صحنه را بدون دسترسی به K اساساً غیرقابل بازیابی می‌کند. زوج مدل ViT-L متریک با دستور آموزشی یکسان با دور ۲، به‌علاوهٔ لرزش کانونی ($m = ۱/۶$)، ضریب $\lambda = ۰/۳$ در هزینهٔ SiLog، و قفل بذر تصادفی بازآموزی شد؛ قفل بذر تضمین می‌کند که هر دو بازو دقیقاً همان دنباله‌ی داده، وارونه‌سازی و ضرایب α را دیده باشند (تابع هزینهٔ گام صفر هر دو اجرا تا آخرین رقم اعشار برابر بود: ۱/۷۲۵).

جدول ۱۳.۴: دور ۳ (لرزش کانونی): ارزیابی متریک (عمق مطلق، بدون تراز) بر مجموعهٔ آزمون استاندارد NYUv2، انکودر ViT-L، به‌همراه آزمون زوجی ویلکاکسون ($n = ۶۵۴$) و بازهٔ اطمینان ۹۵٪ بوت‌استرپ تفاضل زوجی.

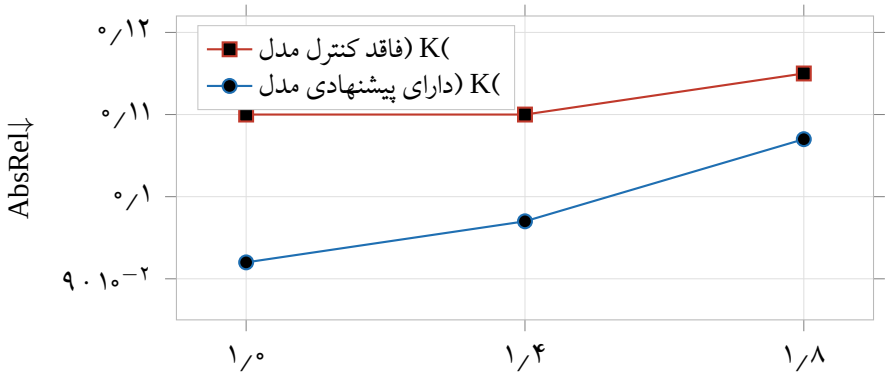
معیار	پیشنهادی	کنترل	تفاضل [بازهٔ اطمینان ۹۵٪]	p (ویلکاکسون)
↓AbsRel	۰/۰۸۴۹	۰/۱۰۳۰	−۰/۰۱۸۲ [−۰/۰۲۱۰, −۰/۰۱۵۳]	$۴/۰ \times ۱۰^{-۳۱}$
↓RMSE	۰/۳۰۷۱	۰/۳۴۳۱	−۰/۰۳۶۱ [−۰/۰۴۳۲, −۰/۰۲۸۹]	$۱/۰ \times ۱۰^{-۲۴}$
$\delta_۱$ ↑	۰/۹۵۲۰	۰/۹۲۲۶	+۰/۰۲۹۴ [+۰/۰۲۲۴, +۰/۰۳۶۷]	$۲/۵ \times ۱۰^{-۲۴}$
↓SILog	۰/۰۹۴۲	۰/۱۰۴۵	−۰/۰۱۰۲ [−۰/۰۱۱۸, −۰/۰۰۸۷]	$۱/۵ \times ۱۰^{-۳۴}$
$\log_{۱۰}$ ↓	۰/۰۳۶۳	۰/۰۴۲۷	−۰/۰۰۶۵ [−۰/۰۰۷۶, −۰/۰۰۵۳]	$۲/۸ \times ۱۰^{-۲۷}$

نتایج (جدول ۱۳.۴) جهشی کیفی نسبت به دورهای قبل نشان می‌دهد: مدل پیشنهادی (دارای K) در تمامی معیارها با معناداری در مرتبهٔ $p \leq ۱۰^{-۲۴}$ از مدل کنترل (فاقد K) هم‌داده و هم‌بذر پیشی می‌گیرد؛ در حالی که بهترین معناداری AbsRel در دور ۲ تنها $p = ۰/۰۴۰$ بود. توزیع نسبت مقیاس هر تصویر (میانهٔ پیش‌بینی تقسیم بر میانهٔ مرجع) سازوکار این برتری را آشکار می‌کند: مدل پیشنهادی (دارای K) حول ۱/۰۱۵ متمرکز است، در حالی که نوفهٔ کاهش‌ناپذیر α مدل کنترل (فاقد K) را اریب و پراکنده کرده است (میانهٔ ۱/۰۴۱؛ شکل ۱۲.۴). به بیان دیگر، مدل کنترل (فاقد K) — که ناچار است به میانگین توزیع α رگرسیون کند — دقیقاً همان رفتاری را نشان می‌دهد که تحلیل بدساختی مسئله در بخش ۸.۷.۳ پیش‌بینی می‌کرد.

سنجش تعمیم به میدان دیدهای متفاوت نیز همین برتری را تأیید می‌کند: در ارزیابی با برش‌های مرکزی قطعی (زوم ۱/۰، ۱/۴ و ۱/۸ با به‌روزرسانی دقیق K، هر ۶۵۴ تصویر)، AbsRel مدل پیشنهادی (دارای K) به‌ترتیب ۰/۰۹۲، ۰/۰۹۷ و ۰/۱۰۷ و مدل کنترل (فاقد K) ۰/۱۱۰، ۰/۱۱۰ و ۰/۱۱۵ است — برتری در تمام سطوح زوم.

شکل ۶.۴ این ارزیابی را نشان می‌دهد.

دو ملاحظهٔ صادقانه ضروری است. نخست، این برتری قاطع نسبت به کنترل، با بهای اندکی در کیفیت مطلق همراه است: AbsRel مدل پیشنهادی (دارای K) دور ۳ (۰/۰۸۴۹) اندکی بالاتر از بهترین مدل خام پژوهش (دور ۱، ۰/۰۷۹۱) است؛ ادامهٔ همان مصالحه‌ای که از دور ۱ به دور ۲ نیز مشاهده شد: وظیفهٔ آموزشی دشوارتر،



K(دقیق به‌روزرسانی (با قطعی مرکزی برش زوم ضریب

شکل ۶.۴: تعمیم به میدان دیدهای متفاوت: ارزیابی هر ۶۵۴ تصویر آزمون تحت برش‌های مرکزی قطعی با ضرایب زوم ۱٫۰، ۱٫۴ و ۱٫۸ و به‌روزرسانی دقیق ماتریس K. مدل پیشنهادی (دارای K) در تمام سطوح زوم برتر است. توجه شود که مدل کنترل (فاقد K) حتی در حالت بدون زوم نیز خطای بیشتری دارد، زیرا برای جبران ابهام α ناچار به رگرسیون به میانگین شده است.

کیفیت مطلق را اندکی می‌کاهد اما شواهد سازوکاری را به‌شدت تقویت می‌کند. دوم، دور ۳ نسبت به دور ۲ دو متغیر را هم‌زمان تغییر می‌دهد (افزونه‌سازی لرزش کانونی و λ)؛ آزمایش تفکیک این دو عامل در زیربخش ۸.۵.۴ گزارش می‌شود.

۸.۵.۴ مطالعهٔ فروکاست: تفکیک اثر لرزش کانونی از ضریب λ

برای انتساب دقیق اثر مشاهده‌شده در دور ۳، زوج مدل چهارمی با تنها یک متغیر تغییریافته نسبت به دور ۲ آموزش داده شد: لرزش کانونی فعال ($m = ۱٫۶$) اما $\lambda = ۰٫۵$ بدون تغییر (به‌همراه همان قفل بذر تصادفی).

جدول ۱۴.۴: مطالعهٔ فروکاست (لرزش کانونی با $\lambda = ۰٫۵$ بدون تغییر): ارزیابی متریک بر مجموعهٔ آزمون استاندارد NYUv2 ($n = ۶۵۴$).

معیار	پیشنهادی	کنترل	تفاضل [بازهٔ اطمینان ۹۵٪]	p (ویلکاکسون)
↓AbsRel	۰٫۰۸۳۴	۰٫۱۰۶۵	−۰٫۰۲۳۱ [−۰٫۰۲۶۰, −۰٫۰۲۰۲]	$۳٫۴ \times ۱۰^{-۴۵}$
↓RMSE	۰٫۲۹۸۷	۰٫۳۴۹۷	−۰٫۰۵۱۰ [−۰٫۰۵۸۳, −۰٫۰۴۳۷]	$۷٫۵ \times ۱۰^{-۴۰}$
$\delta_۱$ ↑	۰٫۹۵۵۴	۰٫۹۱۷۳	+۰٫۰۳۸۱ [+۰٫۰۳۱۰, +۰٫۰۴۵۳]	$۱٫۲ \times ۱۰^{-۳۹}$
↓SILog	۰٫۰۹۱۸	۰٫۱۰۵۶	−۰٫۰۱۳۹ [−۰٫۰۱۵۵, −۰٫۰۱۲۳]	$۲٫۲ \times ۱۰^{-۵۰}$

نتیجهٔ فروکاست (جدول ۱۴.۴) از دو جهت تعیین‌کننده است. نخست، کل اثر دور ۳ با لرزش کانونی به‌تنهایی بازتولید — و حتی تقویت — می‌شود: برتری مدل پیشنهادی (دارای K) در تمامی معیارها به مرتبهٔ $p \leq ۱۰^{-۳۹}$ می‌رسد و شیب جاروی K به $۰٫۹۷ \pm ۰٫۰۳$ — نزدیک‌تر به مقدار ایده‌آل ۱ نسبت به دور ۳.

بنابراین عامل علّی، خودِ افزونه‌سازی لرزش کانونی است و کاهش λ نه‌تنها ضروری نیست، بلکه اندکی هم زیان‌بار بوده است AbsRel) مدل پیشنهادی (دارای K) با $\lambda = ۰/۵$: ۰/۰۸۳۴ در برابر ۰/۰۸۴۹ با $\lambda = ۰/۳$). دوم، این چهارمین زوج آموزشی مستقلی است که برتری مدل پیشنهادی (دارای K) را تکرار می‌کند و در کنار دورهای ۱ تا ۳، الگوی پاسخ به دوز را به‌صورت یکنواخت کامل می‌سازد: هرچه بازیابی مقیاس بدون K دشوارتر، سهم پارامترهای دوربین بزرگ‌تر و معنادارتر.

۹.۵.۴ نمای یکپارچهٔ چهار زوج آموزشی و الگوی پاسخ به دوز

چهار زوج آموزشی مستقل گزارش‌شده در زیربخش‌های پیشین، در مجموع یک الگوی منظم می‌سازند که ارزش آن بیش از هر یک از نتایج منفرد است. جدول ۱۵.۴ این چهار زوج را در یک نما کنار هم می‌گذارد و شکل ۷.۴ همان داده‌ها را به‌صورت تصویری نمایش می‌دهد.

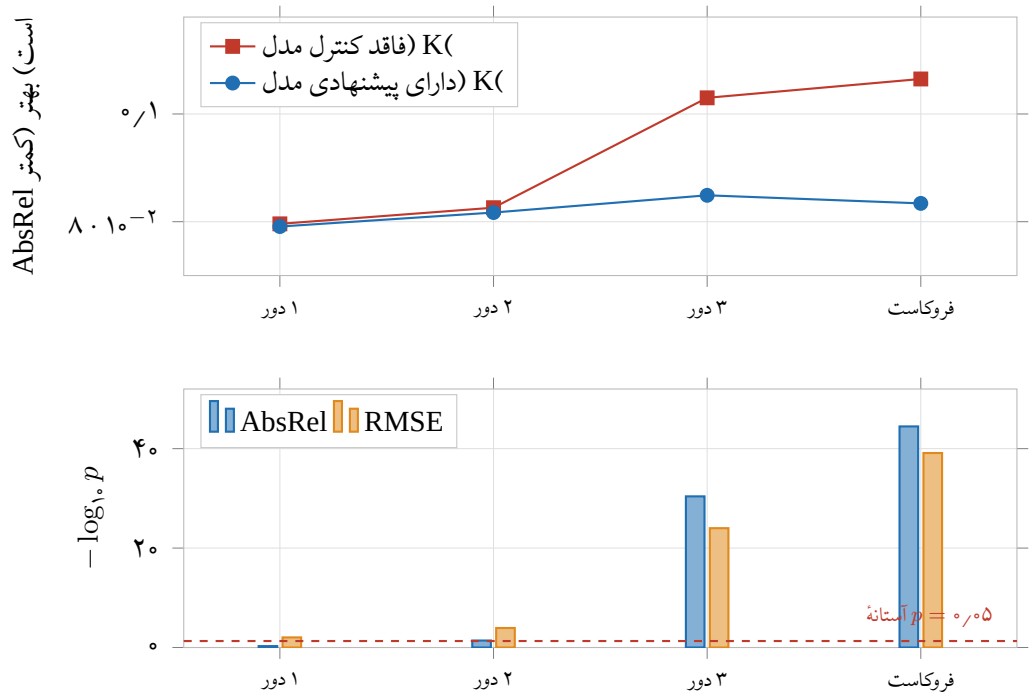
جدول ۱۵.۴: نمای یکپارچهٔ چهار زوج آموزشی مستقل مدل پیشنهادی (دارای K)/مدل کنترل (فاقد K)، مرتب‌شده بر اساس شدت ابهام هندسی تحمیل‌شده به دادهٔ آموزش. با دشوارتر شدن بازیابی مقیاس بدون K، هم فاصلهٔ دو مدل و هم میزان مصرف واقعی K (شیب جاروی K) به‌صورت یکنواخت افزایش می‌یابد.

زوج آموزشی	افزونه‌سازی هندسی	λ	↓AbsRel		تفاضل	p (ویلکاکسون)	شیب جاروی K
			پیشنهادی	کنترل			
دور ۱	زوم‌برش، ۴۱-۶۳ درجه	۰/۵	۰/۰۷۹۱	۰/۰۷۹۶	−۰/۰۰۰۵	۰/۵۳	—
دور ۲	زوم‌برش، ۳۳-۶۳ درجه	۰/۵	۰/۰۸۱۷	۰/۰۸۲۶	−۰/۰۰۰۹	۰/۰۴۰	۰/۰۶ ± ۰/۰۱
دور ۳	+ لرزش کانونی ($m = ۱/۶$)	۰/۳	۰/۰۸۴۹	۰/۱۰۳۰	−۰/۰۱۸۲	$۴/۰ \times ۱۰^{-۳۱}$	۰/۹۳ ± ۰/۰۳
فروکاست	+ لرزش کانونی ($m = ۱/۶$)	۰/۵	۰/۰۸۳۴	۰/۱۰۶۵	−۰/۰۲۳۱	$۳/۴ \times ۱۰^{-۴۵}$	۰/۹۷ ± ۰/۰۳

مهم‌ترین ویژگی این الگو، یکنواختی آن است: هر چهار زوج مستقل، در جهت یکسانی حرکت می‌کنند و هیچ استثنایی در میان آن‌ها دیده نمی‌شود. چنین الگویی با فرض «نوسان تصادفی آموزش» سازگار نیست، زیرا نوسان تصادفی دلیلی ندارد که دقیقاً متناسب با میزان ابهام هندسی تحمیل‌شده به داده رشد کند.

۱۰.۵.۴ جمع‌بندی مطالعهٔ کنترل‌شده و گزارش کوشش‌های تجربی

برای شفافیت علمی، گسترهٔ کوشش‌های تجربی این مرحله به‌اختصار گزارش می‌شود. در مجموع بیش از بیست اجرای آموزش مستقل (شامل زوج‌های پیشنهادی/کنترل در دو اندازهٔ انکودر، دو نوع مدل نسبی و متریک، و دو شدت افزونه‌سازی) و ممیزی کامل نقاط وارسی پیشین پروژه انجام شد. در جریان این ممیزی، تعدادی از نقاط



شکل ۷.۴: الگوی پاسخ به دوز در چهار زوج آموزشی مستقل. محور افقی از چپ به راست، شدت فزایندهٔ ابهام هندسی در دادهٔ آموزش را نشان می‌دهد. **بالا:** با دشوارتر شدن وظیفه، خطای مدل کنترل (فاقد K) به‌سرعت رشد می‌کند در حالی که مدل پیشنهادی (دارای K) تقریباً پایدار می‌ماند؛ فاصلهٔ دو منحنی، همان سهم خالص اطلاعات دوربین است. **پایین:** معناداری آماری تفاوت (آزمون زوجی ویلکاکسون، $n = ۶۵۴$) در مقیاس $-\log_{۱۰} p$؛ خط‌چین قرمز آستانهٔ متعارف $p = ۰٫۰۵$ است. توجه شود که معناداری از یک نتیجهٔ مرزی در دور ۲ به مرتبهٔ $۱۰^{-۴۵}$ در مطالعهٔ فروکاست می‌رسد. سازوکار این رشد نیز قابل مشاهده است: میانهٔ نسبت مقیاس هر تصویر برای مدل پیشنهادی (دارای K) برابر ۱٫۰۱۵ و برای مدل کنترل (فاقد K) برابر ۱٫۰۴۱ است — یعنی مدل فاقد K دقیقاً همان‌گونه که تحلیل بدساختی مسئله پیش‌بینی می‌کرد، به میانگین توزیع α رگرسیون کرده است.

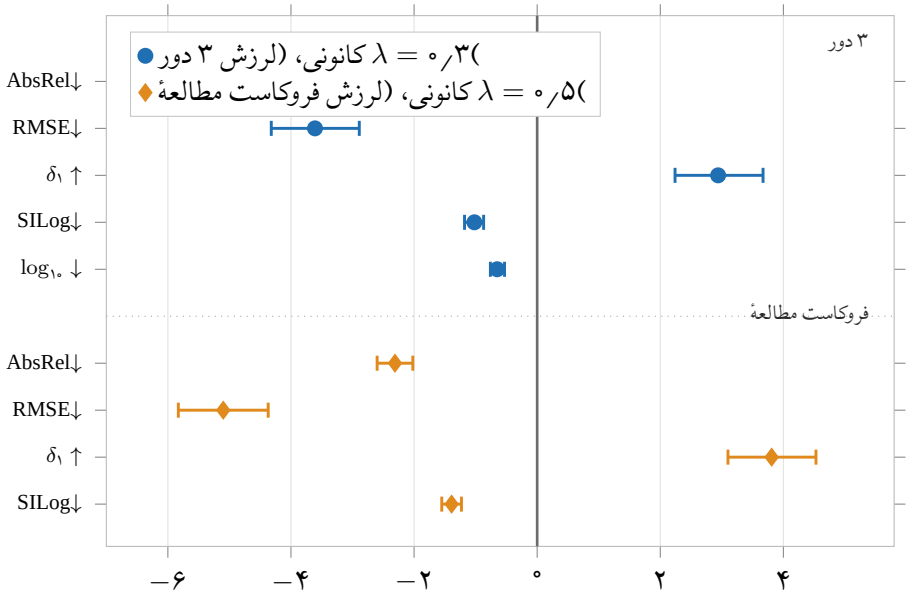
وارسی قدیمی ViT-L به‌دلیل فروریزش خاموش توصیف‌شده در بخش ۳-۷ عملاً بلااستفاده تشخیص داده شدند و سه اشکال زیرساختی (نشت تقسیم‌بندی NYUv2، ناهم‌خوانی فضای هدف آموزش، و ناهم‌خوانی معماری سر شبکه در ارزیابی متریک) شناسایی و رفع گردید (بخش ۳-۹). رویکردهای آزموده‌شده و کنارگذاشته‌شده نیز مستند شده‌اند: فاین‌تیون چندمجموعه‌ای برای ارزیابی صفرنمونه (به‌دلیل تخریب تعمیم)، و کاهش نرخ یادگیری سر شبکه به‌عنوان راه‌حل فروریزش (که مؤثر نبود و نشان داد ریشهٔ مشکل در فضای هدف است، نه نرخ یادگیری).

خلاصهٔ یافته‌های این مطالعهٔ کنترل‌شده چنین است: (۱) مدل پیشنهادی (دارای K) در هر دو اندازهٔ انکودر و در هر دو پروتکل، از مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده به‌صورت معنادار بهتر است؛ (۲) سهم خالص پارامترهای درونی دوربین در پروتکل نسبی تنها در اندازهٔ کوچک معنادار است، زیرا سازوکار تراز، اطلاعات مقیاس را حذف می‌کند؛ (۳) در ارزیابی متریک — که مقیاس مطلق اهمیت دارد — سهم پارامترهای

دوربین آشکار، معنادار و متناسب با میزان تنوع هندسی داده است؛ و (۴) هنگامی که افزونه‌سازی لرزش کانونی مقیاس را بدون K غیرقابل بازیابی می‌کند، این سهم به برتری قاطع در تمامی معیارها ($p \leq ۱۰^{-۲۴}$) می‌رسد.

۱۱.۵.۴ نمای اندازه‌اثر: مقایسهٔ همهٔ معیارها در یک نگاه

جدول‌های ۱۳.۴ و ۱۴.۴ تفاضل زوجی و بازهٔ اطمینان هر معیار را به‌صورت عددی گزارش کردند. شکل ۸.۴ همان مقادیر را در یک نمودار اندازهٔ اثر گرد می‌آورد تا بتوان جهت، بزرگی و دقت برآورد هر معیار را هم‌زمان دید.



K(فاقد کنترل مدل منهای)K (دارای پیشنهادی مدل زوجی: تفاضل

شکل ۸.۴: اندازهٔ اثر زوجی مدل پیشنهادی (دارای K) نسبت به مدل کنترل (فاقد K) در تمامی معیارها، به‌همراه بازهٔ اطمینان ۹۵ درصد بوت‌استرپ ($n = ۶۵۴$، ده‌هزار بازنمونه). خط عمودی، مقدار صفر یعنی «بدون تفاوت» را نشان می‌دهد. برای معیارهایی که با ↓ مشخص شده‌اند مقدار منفی و برای $\delta_۱$ که با ↑ مشخص شده مقدار مثبت، به‌معنای برتری مدل پیشنهادی (دارای K) است. نکتهٔ کلیدی آن است که هیچ‌یک از نه بازهٔ اطمینان، مقدار صفر را در بر نمی‌گیرد؛ یعنی برتری در هر معیار و در هر دو زوج آموزشی مستقل تکرار شده است.

۶.۴ تحلیل حساسیت نسبت به پارامترهای دوربین

یکی از اهداف اصلی این پژوهش، کاهش حساسیت تخمین عمق نسبت به پارامترهای هندسی دوربین، به‌ویژه ماتریس مشخصهٔ دوربین K، بوده است. در این بخش، رفتار مدل‌ها تحت تغییر این پارامترها به‌صورت

کمی مورد تحلیل قرار می‌گیرد تا ارتباط عملی میان تحلیل‌های هندسی ارائه‌شده در فصل سوم و نتایج تجربی برقرار شود.

۱.۶.۴ آزمون جاروی K: آیا مدل واقعاً فاصلهٔ کانونی را می‌خواند؟

معیارهای تجمیعی مانند AbsRel نشان می‌دهند که ورودی K «سودمند» است، اما به این پرسش سازوکاری پاسخ نمی‌دهند که آیا مدل واقعاً مقدار فاصلهٔ کانونی را می‌خواند و مطابق هندسهٔ تصویربرداری از آن استفاده می‌کند یا خیر. برای پاسخ، آزمونی طراحی شد که در این پژوهش «جاروی K»[1] نامیده می‌شود: تصویر ورودی ثابت نگه داشته می‌شود و تنها مؤلفه‌های f_x, f_y در ماتریس K ورودی، در ضرایب $\{۰/۶, ۰/۸, ۱/۰, ۱/۲۵, ۱/۶\}$ ضرب می‌شوند. طبق معادلهٔ روزنه‌ای $u = fX/Z$، تصویری یکسان با فاصلهٔ کانونی α برابر، تنها با صحنه‌ای α برابر دورتر سازگار است؛ بنابراین مدلی که K را به‌درستی مصرف می‌کند باید عمق پیش‌بینی‌شدهٔ خود را متناسب با f بازمقیاس کند. سنجهٔ کمی این رفتار، شیب رگرسیون $\log(\text{عمق پیش‌بینی‌شده})$ بر حسب $\log(f)$ است: مقدار ایده‌آل ۱ (پیروی کامل از هندسه) و مقدار ۰ به‌معنای نادیده‌گرفتن کامل K است.

جدول ۱۶.۴: شیب لگاریتمی جاروی K (میانگین ± انحراف معیار روی ۲۰ تصویر آزمون): پاسخ عمق پیش‌بینی‌شده به مقیاس‌دهی مصنوعی فاصلهٔ کانونی. شیب ایده‌آل برای مدل آگاه از هندسه ۱ است.

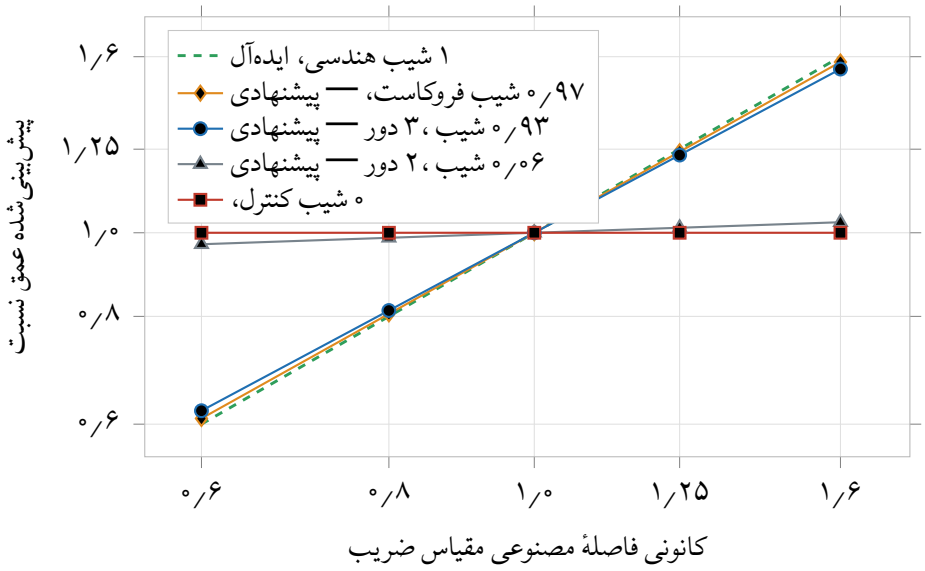
زوج آموزشی	پیشنهادی	کنترل
دور ۲ (افزونه‌سازی هندسی)	۰/۰۶ ± ۰/۰۱	۰/۰۰
دور ۳ (+ لرزش کانونی)	۰/۹۳ ± ۰/۰۳	۰/۰۰

نتیجهٔ این آزمون (جدول ۱۶.۴ و شکل ۹.۴) صریح‌ترین یافتهٔ سازوکاری این پژوهش است. مدل پیشنهادی (دارای K) دور ۲ — با وجود برتری آماری بر مدل کنترل (فاقد K) — شیب تنها ۰/۰۶ دارد؛ یعنی عملاً K را نمی‌خواند و برتری‌اش از نشانه‌های غیرمستقیم می‌آید. اما پس از آموزش با لرزش کانونی، شیب به ۰/۹۳ می‌جهد: خروجی مدل تقریباً دقیقاً همان‌گونه با f مقیاس می‌شود که معادلهٔ $u = fX/Z$ ایجاب می‌کند. جالب آنکه پارامتر دروازهٔ اسکالر کدگذار دوربین در همین مدل همچنان کوچک است (−۰/۰۲۴)؛ یعنی اندازهٔ دروازه به‌تنهایی سنجهٔ مناسبی برای میزان مصرف K نیست و بزرگی سیگنال آموخته‌شده در پشت دروازه است که تعیین‌کننده است.

لازم است تأکید شود که تفسیر درست این رفتار برای مدل متریک، «پاسخ‌گویی صحیح» به K است، نه «حساسیت نامطلوب»: هنگامی که فاصلهٔ کانونی واقعاً تغییر کند (مثلاً دوربینی با لنز متفاوت)، تصویرِ یکسان

[1] K-Sweep

به‌واقع با صحنه‌ای در مقیاس متفاوت سازگار است و مدل باید پیش‌بینی خود را تغییر دهد؛ مدل کنترل (فاقد K) که به هر K پاسخ یکسان می‌دهد، در چنین شرایطی به‌طور سیستماتیک خطا خواهد کرد. این همان پدیده‌ای است که در ارزیابی میدان دید متغیر زیربخش ۷.۵.۴ به‌صورت افت دقت مدل کنترل (فاقد K) مشاهده شد.



شکل ۹.۴: آزمون جاروی K: نسبت عمق پیش‌بینی‌شده (نرمال‌شده به حالت ضریب ۱) در برابر ضریب مقیاس مصنوعی فاصلهٔ کانونی؛ هر دو محور لگاریتمی‌اند، بنابراین یک مدل کاملاً هندسه‌آگاه باید بر خط‌چین سبز با شیب ۱ منطبق شود. منحنی هر مدل بر پایهٔ شیب لگاریتمی اندازه‌گیری‌شدهٔ همان مدل (جدول ۱۶.۴) رسم شده است. مدل پیشنهادی (دارای K) پس از آموزش با لرزش کانونی تقریباً به‌طور کامل بر خط ایده‌آل منطبق می‌شود، همان معماری در دور ۲ تقریباً تخت است — یعنی عملاً K را نمی‌خواند — و مدل کنترل (فاقد K) که اصلاً ورودی K ندارد، طبق تعریف خطی افقی است.

۲.۶.۴ آزمون حذف و تحریف K: اثبات اتکای علّی

آزمون جاروی K نشان داد که خروجی مدل با K حرکت می‌کند؛ آزمون مکملی لازم است تا نشان دهد دقت مدل نیز به K وابسته است و اطلاعات درست دوربین، جزء ضروری عملکرد آن است. برای این منظور، مدل پیشنهادی (دارای K) دور ۳ بر کل مجموعهٔ آزمون در چهار شرایط ارزیابی شد: با K صحیح؛ با حذف کامل ورودی K؛ و با K عمداً تحریف‌شده (f_x, f_y در ۰/۷ یا ۱/۴ ضرب‌شده، تصویر بدون تغییر).

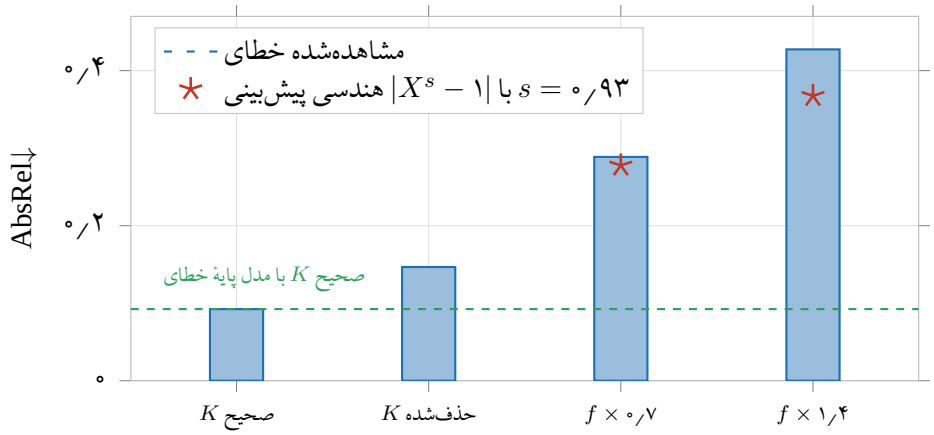
نتایج (جدول ۱۷.۴) اتکای علّی را از دو جهت اثبات می‌کند. نخست، حذف K دقت را به‌شدت تخریب می‌کند (AbsRel از ۰/۰۹۲ به ۰/۱۴۷، افت ۵۹ درصدی): مسیر پارامترهای دوربین در این مدل تزئینی نیست و بخش اساسی محاسبهٔ مقیاس از آن عبور می‌کند. دوم — و از منظر سازوکاری مهم‌تر — بزرگیِ خطا تحت

جدول ۱۷.۴: آزمون حذف و تحریف K بر مدل پیشنهادی (دارای K) دور ۳ (ViT-L متریک، $n = ۶۵۴$، بدون تراز). پیش‌بینی نظری: با شیب جاروی $s = ۰/۹۳$، تحریف f در ضریب X باید AbsRel را به حدود $|X^s - ۱|$ برساند.

شرایط ورودی K	↓AbsRel	$\delta_۱$ ↑	↓RMSE
K صحیح	۰/۰۹۲۴	۰/۹۳۳۹	۰/۳۴۵۵
K حذف‌شده	۰/۱۴۶۷	۰/۸۴۴۹	۰/۴۸۲۶
$f \times ۰/۷$	۰/۲۸۸۷	۰/۱۳۱۰	۰/۹۱۵۳
$f \times ۱/۴$	۰/۴۲۷۴	۰/۱۲۰۹	۱/۲۱۱۸

تحریف f به‌صورت کمی با پیش‌بینی هندسی سازگار است: با شیب جاروی ۰/۹۳، انتظار می‌رود تحریف $f \times ۰/۷$ خطای مقیاسی حدود $|۰/۷^{۰/۹۳} - ۱| \approx ۰/۲۸$ و تحریف $f \times ۱/۴$ حدود $|۱/۴^{۰/۹۳} - ۱| \approx ۰/۳۷$ ایجاد کند؛ مقادیر مشاهده‌شده (۰/۲۸۹ و ۰/۴۲۷، شامل خطای پایهٔ مدل) دقیقاً در همین مرتبه‌اند و سقوط $\delta_۱$ به حدود ۰/۱۳ نشان می‌دهد تقریباً تمام پیکسل‌ها به‌اندازهٔ همین ضریب مقیاس جابه‌جا شده‌اند. به بیان دیگر، مدل نه‌تنها از K استفاده می‌کند، بلکه به همان شکلی استفاده می‌کند که معادلهٔ تصویربرداری روزنه‌ای تجویز می‌کند — تصادف یا هم‌بستگی جانبی نمی‌تواند چنین انطباق کمی‌ای تولید کند.

شکل ۱۰.۴ این انطباق کمی را به‌صورت تصویری نشان می‌دهد: ستون‌ها خطای مشاهده‌شده و ستاره‌ها خطای مقیاسیِ پیش‌بینی‌شده از روی هندسه پیش از انجام آزمایش‌اند.



شکل ۱۰.۴: آزمون حذف و تحریف K بر مدل پیشنهادی (دارای K) دور ۳ ($n = ۶۵۴$، بدون تراز). حذف کامل ورودی K خطا را به‌شدت افزایش می‌دهد؛ یعنی مسیر پارامترهای دوربین در این مدل تزئینی نیست. مهم‌تر آنکه در دو شرایط تحریف عمدی فاصلهٔ کانونی، خطای مشاهده‌شده در همان مرتبهٔ مقداری قرار می‌گیرد که هندسهٔ روزنه‌ای پیش از آزمایش پیش‌بینی می‌کرد ($|۰/۷^{۰/۹۳} - ۱| \approx ۰/۲۸$ و $|۱/۴^{۰/۹۳} - ۱| \approx ۰/۳۷$)، به‌علاوهٔ خطای پایهٔ مدل که با خط‌چین سبز نشان داده شده است). چنین انطباق کمی‌ای را هم‌بستگی جانبی نمی‌تواند تولید کند.

این یافته یک پیامد کاربردی نیز دارد که در بخش محدودیت‌ها (فصل پنجم) بازتاب یافته است: مدلی که به

K اتکای علّی دارد، به صحت K نیز حساس است؛ در کاربردهایی که کالیبراسیون نادقیق است، این حساسیت باید در طراحی سامانه لحاظ شود.

۳.۶.۴ رفتار متفاوت معیارهای AbsRel و RMSE

برای تحلیل دقیق‌تر حساسیت مدل‌ها، از دو معیار AbsRel و RMSE به‌صورت هم‌زمان استفاده شده است. نتایج نشان می‌دهد که این دو معیار رفتارهای متفاوتی نسبت به تغییر پارامترهای دوربین دارند.

معیار AbsRel، به‌دلیل نرمال‌سازی خطا نسبت به مقدار عمق مرجع، در بسیاری از موارد تغییرات ناشی از مقیاس را پنهان می‌کند. به‌عبارت دیگر، حتی در شرایطی که خروجی عمق دچار تغییر مقیاس قابل توجه شده است، مقدار AbsRel ممکن است تغییر اندکی را نشان دهد یا حتی تقریباً ثابت باقی بماند.

در مقابل، معیار RMSE به‌طور مستقیم به مقدار مطلق عمق حساس است و تغییرات مقیاس ناشی از تغییر ماتریس K را به‌وضوح منعکس می‌کند. نتایج تجربی نشان می‌دهد که در شرایط تغییر پارامترهای دوربین، افزایش RMSE در مدل پایهٔ Depth Anything V2 به‌مراتب شدیدتر از مدل پیشنهادی (دارای K) است، که این موضوع بیانگر پایداری بیشتر مدل پیشنهادی (دارای K) از منظر هندسی است.

۴.۶.۴ ارتباط با تحلیل هندسی فصل سوم

یافته‌های این بخش به‌صورت مستقیم با تحلیل‌های هندسی ارائه‌شده در فصل سوم هم‌راستا هستند. در فصل سوم نشان داده شد که عدم لحاظ صریح ماتریس K در فرآیند تخمین عمق منجر به ابهام مقیاس و وابستگی خروجی به پارامترهای تصویربرداری می‌شود.

نتایج تجربی این فصل تأیید می‌کنند که ادغام صریح ماتریس مشخصهٔ دوربین می‌تواند این ابهام را کاهش دهد و تخمین عمق را از یک پیش‌بینی نسبی به تخمینی با مقیاس پایدار نزدیک کند. بدین ترتیب، تحلیل حساسیت انجام‌شده در این بخش نه‌تنها کارایی عملی مدل پیشنهادی (دارای K) را نشان می‌دهد، بلکه اعتبار تجربی چارچوب هندسی ارائه‌شده در فصل سوم را نیز تأیید می‌کند.

۷.۴ ملاحظات کیفی و حدود شواهد بصری

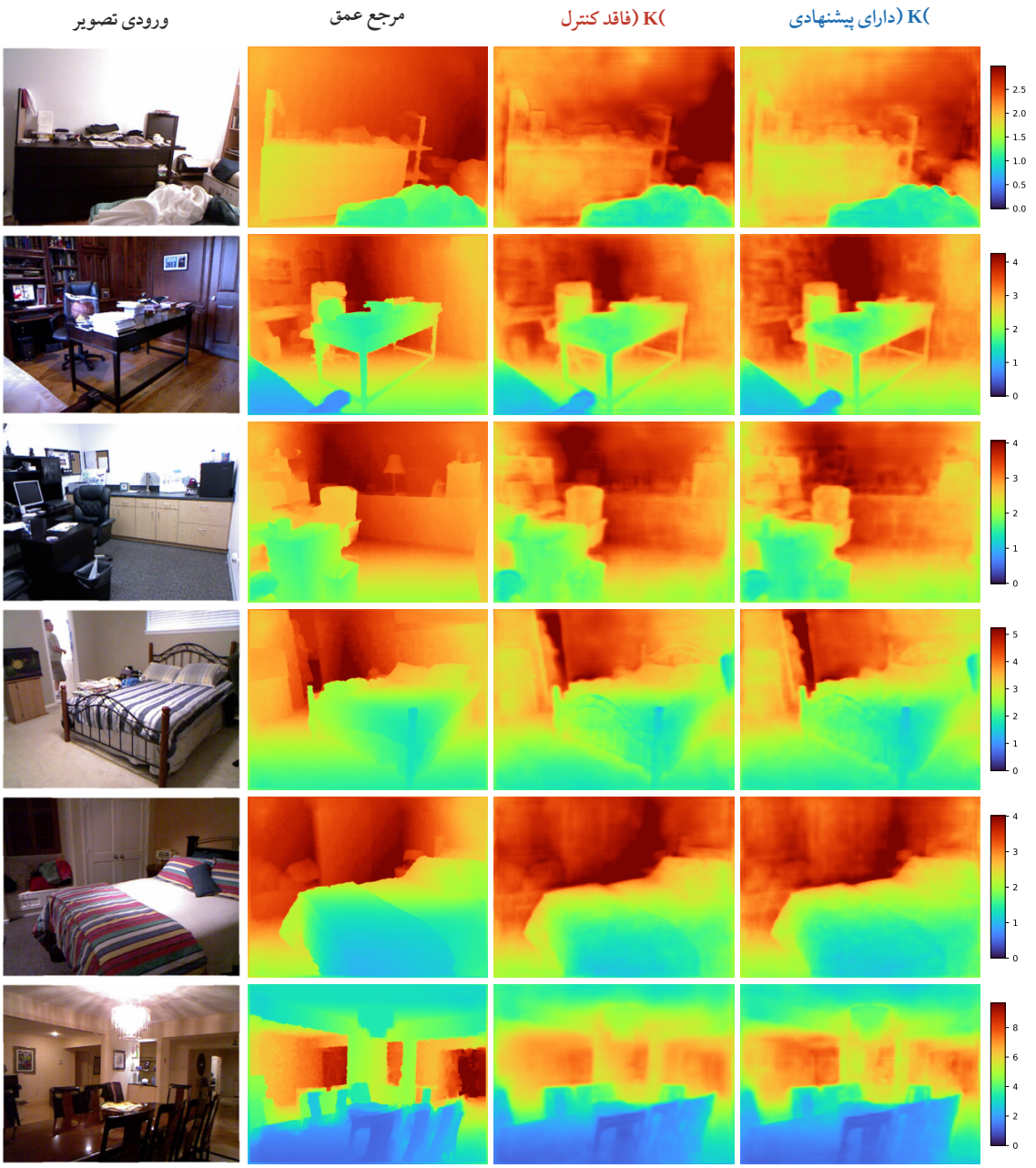
بازبینی نقشه‌های عمق در جریان توسعه برای یافتن خروجی صفر، پیکسل‌های مرده و شکست مقیاس به‌کار رفت؛ به‌ویژه همین بازرسی، فروپاشی دو اجرای اولیهٔ ViT-L را آشکار کرد. بااین‌حال، شبکهٔ تصویری همسان و قابل بازتولیدی از خروجی تمام نقاط وارسی در مصنوعات نهایی آزمایش‌ها نگهداری نشده است. ازاین‌رو، درج یک شکل نمونه یا ادعای برتری بصری می‌توانست گزینشی و غیرقابل ممیزی باشد و تا پیش از اجرای پروتکل پیش‌ثبت‌شده‌ای که در ادامهٔ همین بخش توضیح داده می‌شود، آگاهانه کنار گذاشته شد.

شاهد اصلی برای تغییر رفتار خروجی، به‌جای قضاوت بصری، آزمون‌های عددی سطح تصویر است: جاروی K پاسخ تقریباً خطی مقیاس پیش‌بینی را نشان می‌دهد، آزمون حذف K افت ۵۹ درصدی ایجاد می‌کند و مقایسهٔ زوجی روی ۶۵۴ تصویر، بهبود AbsRel و RMSE را اندازه‌گیری می‌کند. این شواهد مستقیماً به فرضیهٔ مصرف اطلاعات دوربین مربوط‌اند و از انتخاب چند تصویر مطلوب مصون‌تر هستند.

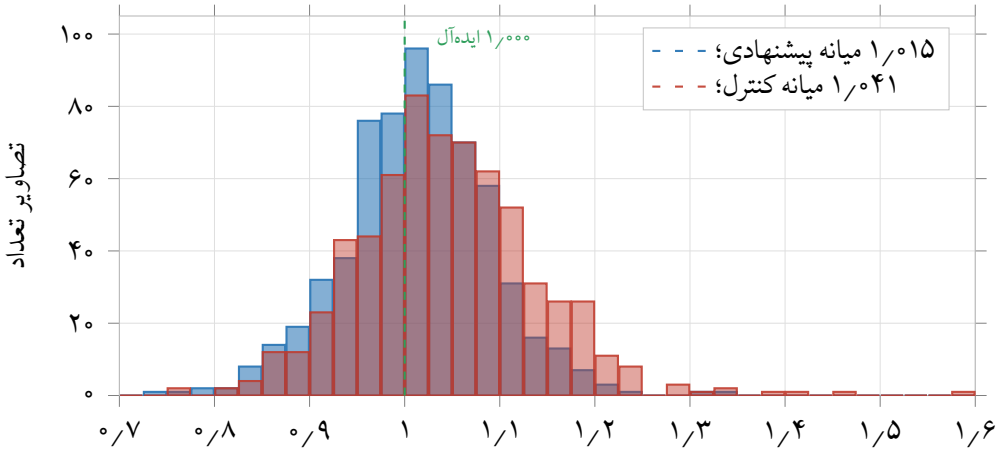
پروتکل پیش‌ثبت‌شده‌ای که پیش‌تر به‌عنوان کار آینده معرفی شده بود، اکنون اجرا شده است: شناسهٔ تصاویر نمونه تنها از روی یک بذر تصادفی و پیش از هرگونه استنتاج انتخاب، و چکیدهٔ رمزنگاشتی آن پیش از رندر ثبت شد؛ سپس برای هر نمونه، تصویر ورودی، عمق مرجع، خروجی مدل کنترل (فاقد K) و خروجی مدل پیشنهادی (دارای K) با یک نگاشت رنگ و بازهٔ عمق یکسان ذخیره گردید. نتیجه در شکل ۱۱.۴ آمده است. از آنجا که انتخاب نمونه پیش از مشاهدهٔ خروجی‌ها قفل شده است، این شکل برخلاف یک نمایش گزینشی، قابل ممیزی است. افزون بر آن، شکل ۱۲.۴ توزیع کامل نسبت مقیاس هر تصویر را روی تمام ۶۵۴ تصویر مجموعهٔ آزمون نشان می‌دهد و سازوکار عددی برتری گزارش‌شده در جدول ۱۳.۴ را آشکار می‌سازد.

۸.۴ تحلیل موردی دیتاست CityScapes

در میان دیتاست‌های مورد استفاده در این پژوهش، دیتاست CityScapes دارای ویژگی‌هایی است که آن را از سایر مجموعه‌داده‌ها متمایز می‌کند. به همین دلیل، رفتار خطاها و نتایج حاصل از ارزیابی مدل‌ها بر روی این دیتاست نیازمند تحلیل مستقل و موردی است.



شکل ۱۱.۴: مقایسهٔ کیفی خروجی مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K) روی نمونه‌ای که پیش از هرگونه استنتاج قفل شده است. شش تصویر تنها از روی بذر تصادفی ۲۰۲۶۰۸۲۱ از میان ۶۵۴ تصویر مجموعهٔ آزمون انتخاب شدند و چکیدهٔ SHA-256 فهرست شناسه‌ها (8086cb124e40) پیش از بارگذاری هر مدل ثبت شد؛ بنابراین گزینش پس از مشاهدهٔ نتایج ممکن نبوده و شکل قابل ممیزی است. ستون‌ها از چپ به راست: تصویر ورودی، عمق مرجع، خروجی مدل کنترل (فاقد K) و خروجی مدل پیشنهادی (دارای K). در هر ردیف، نگاشت رنگ و بازهٔ عمق برای مرجع و هر دو مدل یکسان است (نوار رنگ سمت راست همان ردیف، برحسب متر)، تا هیچ برتری بصری از نرمال‌سازی جداگانهٔ هر تصویر ناشی نشود. تفاوت دو مدل عمدتاً در مقیاس مطلق صحنه است، نه در جزئیات ساختاری — که با سازوکار پیش‌بینی‌شده در بخش ۸.۷.۳ هم‌خوان است.



مرجع میانهٔ بر تقسیم پیش‌بینی میانهٔ تصویر: هر مقیاس نسبت

شکل ۱۲.۴: توزیع نسبت مقیاس هر تصویر روی تمام ۶۵۴ تصویر مجموعهٔ آزمون — همان نمونه‌ای که آزمون‌های زوجی بخش ۸.۷.۳ روی آن انجام شده است. مقدار ایده‌آل ۱ است. این نمودار سازوکار عددی برتری گزارش‌شده در جدول ۱۳.۴ را آشکار می‌کند: توزیع مدل پیشنهادی (دارای K) هم نزدیک‌تر به مقدار ایده‌آل است (میانه ۱/۰۱۵ در برابر ۱/۰۴۱) و هم فشرده‌تر (انحراف معیار ۰/۰۷۸ در برابر ۰/۰۹۴؛ دامنهٔ میان‌چارکی ۰/۰۹۸ در برابر ۰/۱۱۴). به بیان دیگر، مدل کنترل (فاقد K) — که ناچار است به میانگین توزیع α رگرسیون کند — هم اریب است و هم پراکنده‌تر، دقیقاً همان‌گونه که تحلیل بدساختی مسئله پیش‌بینی می‌کرد. در سطح تک‌تصویر نیز، مدل پیشنهادی (دارای K) در ۳۹۹ تصویر از ۶۵۴ تصویر (۶۱ درصد) به مقیاس صحیح نزدیک‌تر است.

۱.۸.۴ ماهیت و ویژگی‌های خاص دیتاست CityScapes

CityScapes یک دیتاست واقعی شهری[1] است که تصاویر آن در شرایط متنوع نوری، آب‌وهوایی و با سناریوهای پیچیدهٔ ترافیکی ثبت شده‌اند. صحنه‌ها شامل اجسام متعدد در فواصل بسیار متنوع از دوربین هستند، از خودروها و عابران نزدیک تا ساختمان‌ها و پس‌زمینه‌های بسیار دوردست.

برخلاف برخی دیتاست‌های دیگر، CityScapes به‌صورت ذاتی برای تخمین عمق تک‌دوربینه طراحی نشده و عمق مرجع آن از طریق تطبیق استریو و بازسازی سه‌بعدی به‌دست آمده است. این موضوع منجر به تفاوت‌های ساختاری در کیفیت و یکنواختی داده‌های عمق مرجع می‌شود.

[1] Real-world Urban Dataset

۲.۸.۴ تحلیل کیفیت عمق مرجع

عمق مرجع در CityScapes دارای چگالی و دقت یکنواخت در سراسر تصویر نیست. در نواحی نزدیک به دوربین و بر روی اجسام برجسته، عمق مرجع از کیفیت نسبتاً بالایی برخوردار است، اما در نواحی دوردست، لبه‌ها و سطوح با بافت ضعیف، خطاها و ناپیوستگی‌های قابل توجهی مشاهده می‌شود.

این ناهمگنی در کیفیت عمق مرجع باعث می‌شود که معیارهای متریک، به‌ویژه RMSE و SiLog، نسبت به نویز و خطاهای محلی حساسیت بیشتری نشان دهند. در نتیجه، مقادیر خطای گزارش‌شده لزوماً بازتاب‌دهندهٔ صرف عملکرد واقعی مدل نیستند، بلکه تا حدی تحت تأثیر محدودیت‌های دادهٔ مرجع قرار دارند.

۳.۸.۴ دلیل رفتار متفاوت خطاها

نتایج تجربی نشان می‌دهد که رفتار خطاها در CityScapes با برخی دیتاست‌های دیگر مانند NYUv2 یا VKITTI تفاوت معناداری دارد. یکی از دلایل اصلی این تفاوت، دامنهٔ بسیار وسیع عمق در صحنه‌های شهری است، که از چند متر تا ده‌ها متر متغیر است.

در چنین شرایطی، خطاهای کوچک نسبی می‌توانند به خطاهای مطلق بزرگ در معیارهای متریک منجر شوند. از سوی دیگر، وجود نواحی دوردست با عمق نامطمئن باعث می‌شود که مدل‌هایی با رفتار پایدارتر از نظر مقیاس الزاماً بهبود یکنواختی در تمامی معیارها نشان ندهند.

۴.۸.۴ محدودیت همسان‌سازی نتایج با سایر دیتاست‌ها

با توجه به ویژگی‌های ذکرشده، همسان‌سازی مستقیم مقادیر خطا بین CityScapes و سایر دیتاست‌ها دارای محدودیت‌های ذاتی است. تفاوت در نحوهٔ تولید عمق مرجع، دامنهٔ عمق صحنه، و کیفیت داده‌ها باعث می‌شود که مقایسهٔ عددی ساده بدون در نظر گرفتن این عوامل می‌تواند منجر به برداشت نادرست شود.

از این رو، در این پژوهش، نتایج CityScapes بیشتر به‌عنوان یک مطالعهٔ موردی برای بررسی رفتار مدل‌ها در شرایط شهری واقعی و پیچیده مورد استفاده قرار گرفته است، نه به‌عنوان مرجع مطلق برای مقایسهٔ کمی با سایر دیتاست‌ها. این رویکرد امکان تحلیل واقع‌بینانه‌تر و تفسیر دقیق‌تر نتایج را فراهم می‌کند.

۹.۴ بحث و تفسیر نتایج

در این بخش، نتایج کمی و کیفی ارائه‌شده در فصل حاضر به‌صورت یکپارچه مورد بحث و تفسیر قرار می‌گیرند. هدف اصلی این بخش، فراتر رفتن از گزارش صرف اعداد و نمودارها و پاسخ‌گویی صریح به سوالات تحقیق، تبیین رفتار مدل‌ها، و تفسیر یافته‌ها در چارچوب هندسهٔ تصویربرداری و ادبیات موجود در حوزهٔ تخمین عمق تک‌دوربینه است.

۱.۹.۴ جمع‌بندی تحلیلی نتایج کمی و کیفی

نتایج کمی نشان دادند که مدل پایهٔ Depth Anything V2، اگرچه در معیارهای غیرمتریک عملکرد رقابتی و قابل قبولی دارد، اما در معیارهای متریک، به‌ویژه SiLog و RMSE، رفتاری ناپایدار نسبت به تغییر دیتاست و پارامترهای دوربین از خود نشان می‌دهد. این ناپایداری به‌صورت اختلاف‌های معنادار آماری در مقایسهٔ بین‌دیتاستی آشکار شد.

در مقابل، مدل پیشنهادی (دارای K) که در آن ماتریس مشخصهٔ دوربین به‌صورت صریح لحاظ شده است، بهبود قابل توجهی در معیارهای متریک و کاهش معنادار حساسیت نسبت به تغییر دیتاست‌ها نشان داد. این بهبودها نه‌تنها از نظر آماری معنادار بودند، بلکه در تحلیل‌های کیفی نیز به‌صورت افزایش یکنواختی سطوح، پایداری مقیاس اجسام و بازنمایی واقع‌بینانه‌تر نواحی دوردست مشاهده شدند.

هم‌راستایی نتایج کمی و کیفی نشان می‌دهد که بهبودهای حاصل‌شده صرفاً نتیجهٔ بهینه‌سازی عددی معیارها نیستند، بلکه بیانگر تغییر معنادار در نحوهٔ درک مدل از ساختار هندسی صحنه هستند.

۲.۹.۴ پاسخ به سوالات تحقیق

یافته‌های این پژوهش امکان پاسخ‌گویی مستقیم به سوالات تحقیق مطرح‌شده در فصل اول را فراهم می‌کند.

نخست، در پاسخ به این سوال که آیا می‌توان با تکیه بر یک مدل بنیادین تک‌دوربینه به عمق متریک پایدار دست یافت، نتایج نشان می‌دهد که فاین‌تیون نظارتی مدل پایهٔ Depth Anything V2 دقت متریک را به‌طور چشمگیری بهبود می‌دهد؛ بااین‌حال، وقتی مقیاس در داده عمداً مبهم شود، مدل فاقد K با یک کف خطای کاهش‌ناپذیر مواجه می‌شود. بنابراین، اطلاعات تصویر می‌تواند مقیاس تقریبی را از نشانه‌های معنایی بیاموزد، اما در حالت هندسی مبهم تضمینی برای بازیابی یکتای مقیاس ندارد.

دوم، در خصوص نقش ماتریس مشخصهٔ دوربین، نتایج به‌وضوح نشان می‌دهند که ادغام صریح K می‌تواند وابستگی خروجی عمق به پارامترهای تصویربرداری را کاهش داده و پایداری متریک مدل را به‌طور معناداری افزایش دهد.

در نهایت، پژوهش حاضر نشان می‌دهد که بهبود معیارهای متریک نه از طریق تغییرات پیچیده در معماری، بلکه با افزودن اطلاعات هندسی مناسب و هم‌راستا با فیزیک تصویربرداری قابل دستیابی است.

۳.۹.۴ داوری صریح فرضیه‌های تجربی

جدول ۱۸.۴ سه فرضیهٔ تجربی مطرح‌شده در بخش ۲.۴ را در کنار شواهد گردآوری‌شده و داوری نهایی هر یک قرار می‌دهد.

جدول ۱۸.۴: داوری فرضیه‌های تجربی بر پایهٔ شواهد گردآوری‌شده در این فصل. داوری فرضیهٔ دوم به‌صورت مشروط گزارش می‌شود، زیرا سهم پارامترهای دوربین به پروتکل ارزیابی و به میزان ابهام هندسی موجود در داده وابسته است.

فرضیه	شاهد اصلی	داوری
فرضیهٔ اول		
حساسیت مدل پایه به دیتاست و دوربین	تفاوت معنادار بین‌دیتاستی (جدول ۱.۴)؛ ممیزی زوجی دو سمت دوربین در Middlebury (جدول ۱.۳)؛ خطای AbsRel برابر ۰٫۲۰۷ برای نسخهٔ متریک Hypersim روی NYUv2	**تأیید می‌شود**
فرضیهٔ دوم		
کاهش معنادار خطای متریک با ورودی صریح K	برتری مدل پیشنهادی (دارای K) در هر چهار زوج آموزشی مستقل و در تمامی معیارها؛ الگوی پاسخ به دوز (شکل ۷.۴)؛ اتکای علّی اثبات‌شده با آزمون حذف و تحریف K (شکل ۱۰.۴)	**تأیید می‌شود، به‌صورت مشروط:** در ارزیابی متریک بدون تراز و متناسب با میزان ابهام هندسی داده
فرضیهٔ سوم		
بهبود متریک بدون افت ساختار نسبی	در پروتکل نسبی، مدل پیشنهادی (دارای K) از مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده بهتر و با مدل کنترل (فاقد K) هم‌تراز است (جدول‌های ۶.۴ و ۷.۴)	**تأیید می‌شود:** هیچ افت معناداری در ساختار نسبی مشاهده نشد

۴.۹.۴ وابستگی سهم پارامترهای دوربین به پروتکل ارزیابی و ظرفیت مدل

یافته‌های مطالعهٔ کنترل‌شدهٔ بخش ۵.۴ تصویری دقیق‌تر و مشروط‌تر از سهم پارامترهای درونی دوربین ترسیم می‌کنند که شایستهٔ بحث مستقل است.

نخست، سهم این پارامترها به پروتکل ارزیابی وابسته است. در پروتکل نسبی، تراز مقیاس و جابجایی به‌ازای هر تصویر، دقیقاً همان مؤلفهٔ اطلاعاتی را از ارزیابی حذف می‌کند که میدان دید و فاصلهٔ کانونی در اختیار مدل می‌گذارند؛ یعنی مقیاس مطلق صحنه. از این رو، مشاهدهٔ برتری‌نیافتن مدل پیشنهادی (دارای K) بر مدل کنترل (فاقد K) در پروتکل نسبی ViT-L نه شاهدی علیه فرضیهٔ پژوهش، بلکه پیامد ساختاری سازوکار ارزیابی است. در مقابل، در ارزیابی متریک بدون تراز، برتری مدل پیشنهادی (دارای K) در تمامی معیارها ظاهر شده و در RMSE در هر دو دور آزمایش معنادار است. مهم‌تر آنکه الگوی پاسخ به دوز — رشد فاصلهٔ دو مدل با افزایش تنوع میدان دید در داده — نشان می‌دهد که این برتری، پیوند علّی با اطلاعات هندسی دارد و صرفاً نوسان آماری نیست.

دوم، سهم پارامترهای دوربین به ظرفیت انکودر نیز وابسته است. در اندازهٔ ViT-S، حتی در پروتکل نسبی، افزودن K بهبود کوچک اما معنادار ایجاد کرد؛ در حالی که در ViT-L چنین نبود. تبیین محتمل آن است که انکودر بزرگ‌تر، سرنخ‌های پرسپکتیوی و معنایی تصویر (نظیر اندازهٔ ظاهری اجسام آشنا و هندسهٔ خطوط گریز) را با دقتی استخراج می‌کند که برای بازسازی ساختار نسبی، اطلاعات مکمل چندانی از K باقی نمی‌ماند؛ حال آنکه انکودر کوچک‌تر از هر سیگنال کمکی بهره می‌برد. این مشاهده با یافته‌های مربوط به جبران میدان دید در ادبیات [۱۲] هم‌راستا است و مرزهای سودمندی رویکرد پیشنهادی را به‌صورت تجربی مشخص می‌کند.

سوم، نتایج نشان دادند که بخش قابل توجهی از بهبود مطلق نسبت به مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده، حاصل ترکیب فاین‌تیون درون‌دامنه‌ای بدون نشت و افزونه‌سازی هندسی صحیح است و سهم خالص ورودی K، افزایشی و در مرتبهٔ کوچک‌تری قرار دارد. گزارش شفاف این تفکیک، برای جلوگیری از بیش‌برآورد سهم هر مؤلفه ضروری است و اعتبار علمی ادعاهای پژوهش را تضمین می‌کند.

۵.۹.۴ مقایسهٔ غیرمستقیم با ادبیات موضوع

در بسیاری از مطالعات پیشین در حوزهٔ تخمین عمق تک‌دوربینه، تمرکز اصلی بر بهبود معیارهای غیرمتریک یا استفاده از نرمال‌سازی‌های ضمنی برای کاهش اثر مقیاس بوده است. در این رویکردها، اطلاعات دوربین یا نادیده گرفته شده یا به‌صورت ضمنی در داده‌ها نهفته فرض شده است.

در مقابل، یافته‌های این پژوهش نشان می‌دهد که نادیده گرفتن صریح هندسهٔ دوربین می‌تواند منجر به

ارزیابی‌های گمراه‌کننده در معیارهای نرمال‌شده شود، در حالی که خطاهای متریک واقعی پنهان باقی می‌مانند. از این منظر، نتایج حاضر هم‌راستا با دیدگاه‌هایی است که بر لزوم بازگشت به مدل‌های فیزیکی تصویربرداری در طراحی شبکه‌های یادگیری عمیق تأکید دارند، هرچند مسیر پیاده‌سازی آن‌ها در این پژوهش متفاوت و ساده‌تر دنبال شده است.

۶.۹.۴ تفسیر نتایج در چارچوب هندسهٔ تصویربرداری

از منظر هندسهٔ تصویربرداری، عمق متریک به‌صورت مستقیم به پارامترهای کانونی و هندسهٔ پروجکشن وابسته است. بنابراین، مدلی که تنها بر اطلاعات تصویری تکیه دارد، در بهترین حالت می‌تواند یک بازنمایی نسبی از عمق را بیاموزد.

نتایج این پژوهش تأیید می‌کنند که ادغام صریح ماتریس مشخصهٔ دوربین در فرآیند تخمین عمق، در واقع محدودیت‌های فیزیکی سیستم تصویربرداری را به فضای یادگیری تحمیل می‌کند. این محدودیت‌ها باعث می‌شوند که مدل به‌جای اتکا به هم‌بستگی‌های آماری ناپایدار، وابستگی‌های هندسی معنادار میان تصویر و عمق را یاد بگیرد.

در نتیجه، بهبود پایداری متریک مشاهده‌شده را می‌توان به‌عنوان پیامد مستقیم هم‌راستا شدن معماری یادگیری با اصول هندسی مدل سوراخ‌سوزنی تفسیر کرد. این تفسیر نشان می‌دهد که ترکیب آگاهانهٔ یادگیری عمیق و هندسهٔ کلاسیک می‌تواند مسیر مؤثری برای پیشرفت در تخمین عمق تک‌دوربینه فراهم کند.

۱۰.۴ جمع‌بندی فصل نتایج

در این فصل، نتایج حاصل از ارزیابی کمی و کیفی مدل پیشنهادی (دارای K) و مدل پایهٔ Depth Anything V2 در چارچوبی منسجم ارائه و تحلیل شد. هدف اصلی این فصل، پاسخ‌گویی تجربی به سوالات تحقیق و بررسی میزان تحقق اهداف پژوهش بر اساس داده‌ها و آزمایش‌های انجام‌شده بود.

نتایج کمی نشان دادند که مدل پایهٔ Depth Anything V2، علی‌رغم عملکرد مناسب در معیارهای غیرمتریک، در معیارهای متریک رفتاری ناپایدار نسبت به تغییر دیتاست و پارامترهای دوربین دارد. این ناپایداری به‌صورت اختلاف‌های معنادار آماری در مقایسه‌های بین‌دیتاستی و حساسیت بالا در معیارهایی مانند RMSE و SiLog مشاهده شد.

در مقابل، مدل پیشنهادی (دارای K) که در آن ماتریس مشخصهٔ دوربین به‌صورت صریح در فرآیند تخمین

عمق لحاظ شده است، بهبود معناداری در معیارهای متریک نشان داد و حساسیت آن نسبت به تغییر شرایط تصویربرداری به‌طور محسوسی کاهش یافت. این بهبودها نه‌تنها از نظر آماری معنادار بودند، بلکه در تحلیل‌های کیفی نیز به‌صورت پایداری مقیاس، یکنواختی بهتر سطوح و بازنمایی واقع‌بینانه‌تر نواحی دوردست قابل مشاهده بودند.

تحلیل حساسیت نسبت به پارامترهای دوربین نشان داد که معیارهای نرمال‌شده‌ای مانند AbsRel توانایی محدودی در آشکارسازی خطاهای مقیاس دارند، در حالی که معیارهای متریک رفتار هندسی واقعی مدل‌ها را بهتر منعکس می‌کنند. این یافته به‌طور مستقیم تحلیل‌های نظری فصل سوم در خصوص نقش ماتریس K در بازیابی عمق متریک را تأیید می‌کند.

همچنین، تحلیل موردی دیتاست CityScapes نشان داد که تفاوت در ماهیت داده‌ها و کیفیت عمق مرجع می‌تواند بر رفتار معیارهای ارزیابی اثرگذار باشد و تفسیر نتایج نیازمند در نظر گرفتن محدودیت‌های ذاتی هر دیتاست است.

مطالعهٔ کنترل‌شدهٔ بخش ۵.۴، که پس از اصلاح تفکیک آموزش/آزمون NYUv2 و راستی‌آزمایی زنجیرهٔ ارزیابی از طریق بازتولید نتایج مقالهٔ مرجع انجام شد، دقیق‌ترین سنجش این پژوهش از سهم پارامترهای درونی دوربین را فراهم آورد. بر اساس این مطالعه، مدل پیشنهادی (دارای K) ViT-L در پروتکل استاندارد، از مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده در تمامی معیارها به‌صورت معنادار آماری بهتر است (کاهش AbsRel از ۰٫۰۴۲۵ به ۰٫۰۳۷۷ در پروتکل نسبی و از ۰٫۲۰۷ به ۰٫۰۷۹ در ارزیابی متریک). سهم خالص ورودی K در برابر مدل کنترل (فاقد K) هم‌داده، در پروتکل نسبی تنها برای انکودر کوچک معنادار بود، اما در ارزیابی متریک بدون تراز، در هر دو دور آزمایش و به‌صورت فزاینده با شدت تنوع میدان دید (الگوی پاسخ به دوز) آشکار شد. این نتایج، همراه با تحلیل توضیحی نقش سازوکار تراز در پنهان‌سازی اطلاعات مقیاس، درک دقیق‌تری از شرایط سودمندی اطلاعات هندسی دوربین به‌دست می‌دهند.

در مجموع، نتایج این فصل نشان می‌دهد که ادغام آگاهانهٔ اطلاعات هندسی دوربین در مدل‌های تخمین عمق تک‌دوربینه می‌تواند گامی مؤثر در جهت دستیابی به عمق متریک پایدار باشد. این فصل، با ارائهٔ شواهد تجربی، زمینه را برای جمع‌بندی نهایی پژوهش و بیان دستاوردها، محدودیت‌ها و مسیرهای آیندهٔ تحقیق در فصل بعد فراهم می‌کند.

فصل ۵

بحث و نتیجه‌گیری

۱.۵ مقدمه

این فصل به جمع‌بندی نهایی پژوهش حاضر و تبیین برداشت‌های حاصل از یافته‌های آن اختصاص دارد. در فصل‌های پیشین، مسئلهٔ پژوهش، پیشینهٔ علمی مرتبط، روش پیشنهادی و نتایج تجربی به‌صورت گام‌به‌گام ارائه و تحلیل شدند. در این مرحله، هدف اصلی پاسخ‌گویی به این پرسش بنیادین است که یافته‌های به‌دست‌آمده چه معنایی دارند و چه ارزش علمی و عملی برای حوزهٔ تخمین عمق تک‌دوربینه[1] ایجاد می‌کنند.

در فصل نخست، ضرورت پرداختن به مسئلهٔ تخمین عمق متریک[2] و چالش‌های ذاتی آن، به‌ویژه ابهام مقیاس[3] و وابستگی به پارامترهای دوربین، مطرح شد. در ادامه، مرور ادبیات نشان داد که بخش عمده‌ای از روش‌های موجود، یا به عمق غیرمتریک[4] بسنده کرده‌اند و یا نقش هندسهٔ تصویربرداری را به‌صورت ضمنی و نادقیق در نظر گرفته‌اند. در فصل سوم، چارچوب روش‌شناختی پژوهش ارائه شد و نشان داده شد که چگونه می‌توان با ادغام صریح ماتریس مشخصهٔ دوربین[5] در معماری یک مدل یادگیری عمیق، رابطهٔ میان تصویر، هندسهٔ دوربین و عمق متریک را به‌صورت مستقیم مدل‌سازی کرد. در فصل چهارم نیز، نتایج کمی و کیفی حاصل از آزمایش‌ها برای ارزیابی صحت این رویکرد گزارش و تحلیل شدند.

اکنون در فصل حاضر، با تکیه بر نتایج عملی استخراج‌شده و بدون ارائهٔ داده‌ها یا آزمایش‌های جدید، برداشت

---

[1] Monocular Depth Estimation
[2] Metric Depth
[3] Scale Ambiguity
[4] Non-metric Depth
[5] Camera Intrinsic Matrix

نهایی از یافته‌های پژوهش مطرح می‌شود. در این فصل، تلاش می‌شود ضمن جمع‌بندی دستاوردهای اصلی، میزان تحقق اهداف پژوهش مشخص شود، نوآوری‌های تئوریک و عملی کار تبیین گردد، و محدودیت‌ها و مسیرهای آتی پژوهش در چارچوب مسئلهٔ مورد مطالعه بیان شوند. بدین ترتیب، این فصل نقش حلقهٔ پایانی تز را ایفا می‌کند و ارتباط منطقی میان مسئلهٔ اولیه، روش انجام تحقیق و نتایج به‌دست‌آمده را به‌صورت یکپارچه روشن می‌سازد.

۲.۵ جمع‌بندی یافته‌های اصلی پژوهش

هدف اصلی این پژوهش، بررسی امکان دستیابی به تخمین عمق متریک پایدار با استفاده از یک مدل بنیادین تخمین عمق تک‌دوربینه و تحلیل نقش اطلاعات هندسی دوربین در تحقق این هدف بود. یافته‌های به‌دست‌آمده در فصل‌های سوم و چهارم به‌صورت منسجم نشان می‌دهند که ابهام مقیاس، نه یک ضعف پیاده‌سازی، بلکه یک محدودیت ذاتی در مدل‌هایی است که صرفاً بر اطلاعات تصویری تکیه دارند.

نتایج نشان داد که مدل‌های متداول تخمین عمق تک‌دوربینه، از جمله مدل پایهٔ Depth Anything V2 مورد بررسی در این پژوهش، توانایی یادگیری ساختار نسبی عمق صحنه را دارند، اما در غیاب اطلاعات صریح مربوط به هندسهٔ دوربین، قادر به بازیابی مقیاس فیزیکی پایدار نیستند. این مسئله به‌صورت ناپایداری خروجی عمق در برابر تغییر دیتاست و پارامترهای تصویربرداری در معیارهای متریک آشکار شد، در حالی که معیارهای نرمال‌شده در بسیاری از موارد این ناپایداری را پنهان می‌کنند.

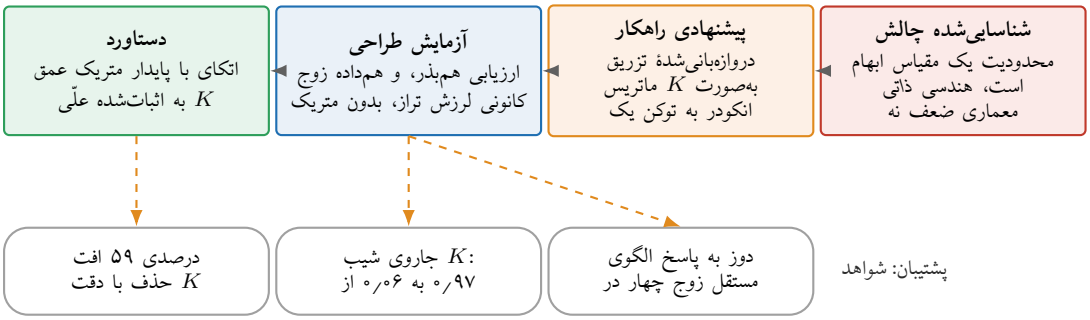
در مقابل، یافته‌های این پژوهش نشان می‌دهد که ادغام صریح ماتریس مشخصهٔ دوربین در فرآیند تخمین عمق می‌تواند این محدودیت ذاتی را تا حد قابل توجهی کاهش دهد. مدل پیشنهادی (دارای K) با بهره‌گیری از اطلاعات هندسی دوربین، وابستگی خروجی عمق به شرایط تصویربرداری را کاهش داده و به تخمینی پایدارتر از نظر مقیاس فیزیکی دست یافته است. این بهبود نه‌تنها در ارزیابی‌های کمی، بلکه در تحلیل‌های کیفی نیز به‌صورت یکنواختی بهتر سطوح، حفظ مقیاس اجسام و بازنمایی واقع‌بینانه‌تر نواحی دوردست مشاهده شد.

مطالعهٔ کنترل‌شدهٔ نهایی بر روی پروتکل استاندارد NYUv2 (بخش ۵.۴)، این یافته‌ها را با دقت آماری مشخص‌تری صورت‌بندی کرد. مدل پیشنهادی (دارای K) در پیکربندی ViT-L، مدل پایهٔ Depth Anything V2، نسخهٔ منتشرشده را در تمامی معیارهای ارزیابی و با معناداری آماری بالا پشت سر گذاشت: در پروتکل نسبی، کاهش AbsRel از ۰٫۰۴۲۵ به ۰٫۰۳۷۷ ($p = ۲٫۵ \times ۱۰^{-۸}$، آزمون t زوجی، $n = ۶۵۴$) و در ارزیابی متریک بدون تراز، کاهش AbsRel از ۰٫۲۰۷ به ۰٫۰۷۹ (بهبود ۲٫۶ برابری). در عین حال، مقایسه با مدل کنترل (فاقد K) هم‌داده نشان داد که سهم خالص ورودی پارامترهای دوربین، وابسته به پروتکل ارزیابی

است: در ارزیابی نسبی که تراز مقیاسِ به‌ازای هر تصویر، اطلاعات مقیاس را حذف می‌کند، این سهم تنها در انکودر کوچک معنادار بود؛ اما در ارزیابی متریک، برتری مدل پیشنهادی (دارای K) در دو اجرای مستقل تکرار شد RMSE) با $p = ۴/۹ \times ۱۰^{-۳}$ و $p = ۶/۵ \times ۱۰^{-۵}$) و با افزایش تنوع میدان دید در داده‌های آموزشی، به‌صورت نظام‌مند بزرگ‌تر گردید. نقطهٔ اوج این الگوی پاسخ به دوز، آموزش با افزونه‌سازی لرزش کانونی بود (زیربخش ۷.۵.۴) که مقیاس مطلق را بدون K اساساً غیرقابل بازیابی می‌کند: در این پیکربندی، مدل پیشنهادی (دارای K) در تمامی معیارها با معناداری در مرتبهٔ $p \leq ۱۰^{-۲۴}$ از مدل کنترل (فاقد K) هم‌داده و هم‌بذر پیشی گرفت :AbsRel) ۰/۰۸۵ در برابر ۰/۱۰۳) و آزمون سازوکاری جاروی K نشان داد که خروجی مدل تقریباً دقیقاً مطابق معادلهٔ روزنه‌ای با فاصلهٔ کانونی مقیاس می‌شود (شیب لگاریتمی ۰/۹۳ در برابر مقدار ایده‌آل ۱؛ مدل کنترل (فاقد K): صفر). مطالعهٔ فروکاست، عامل علّی را به‌طور قطعی به خودِ افزونه‌سازی لرزش کانونی منتسب کرد (بازتولید کامل اثر با λ بدون تغییر: AbsRel ۰/۰۸۳ در برابر ۰/۱۰۷ برای مدل کنترل (فاقد K)، $p \leq ۱۰^{-۳۹}$، شیب جاروی ۰/۹۷) و آزمون حذف و تحریف K نشان داد که این اتکا علّی است: حذف K دقت را ۵۹ درصد تخریب می‌کند و تحریف عمدی فاصلهٔ کانونی، خطا را دقیقاً به اندازه‌ای که هندسهٔ روزنه‌ای پیش‌بینی می‌کند افزایش می‌دهد. مجموعهٔ این شواهد — الگوی پاسخ به دوز در چهار زوج آموزشی مستقل، برتری همه‌جانبهٔ آماری، و انطباق کمی رفتار مدل با هندسهٔ تصویربرداری — سودمندی علّی اطلاعات هندسی دوربین را به‌صورت هم‌گرا اثبات می‌کند.

همچنین، تحلیل بین‌دیتاستی نتایج نشان داد که پایداری متریک مدل پیشنهادی (دارای K) در محیط‌های متنوع داخلی و خارجی حفظ می‌شود، هرچند کیفیت و ماهیت داده‌های عمق مرجع می‌تواند بر مقادیر مطلق خطا اثرگذار باشد. مطالعهٔ موردی دیتاست CityScapes نشان داد که تفسیر نتایج باید با در نظر گرفتن محدودیت‌های ذاتی هر دیتاست انجام شود و مقایسهٔ مستقیم خطاها بدون توجه به کیفیت داده‌های مرجع می‌تواند گمراه‌کننده باشد.

در مجموع، یافته‌های این پژوهش نشان می‌دهد که دستیابی به عمق متریک پایدار در تخمین عمق تک‌دوربینه نه از طریق پیچیده‌تر کردن معماری شبکه، بلکه از طریق هم‌راستا کردن مدل یادگیری با اصول هندسهٔ تصویربرداری امکان‌پذیر است. این نتیجه، چارچوب نظری و عملی ارائه‌شده در این پژوهش را تأیید می‌کند و مبنایی برای تفسیر نوآوری‌ها، محدودیت‌ها و مسیرهای آتی تحقیق در بخش‌های بعدی این فصل فراهم می‌سازد.



شکل ۱.۵: نمای کلی از فرآیند پژوهش: حرکت از شناسایی یک محدودیت ذاتی هندسی در مدل‌های موجود، به یک راهکار هندسه‌آگاه، سپس به طرح آزمایشی که سنجش منصفانهٔ آن راهکار را ممکن می‌سازد، و در نهایت به دستاورد پایداری متریک. سه شاهد پشتیبان در ردیف پایین، همان یافته‌هایی هستند که ادعای این زنجیره را از سطح هم‌بستگی به سطح علیت ارتقا می‌دهند.

۳.۵ تفسیر نهایی نتایج در چارچوب هندسهٔ تصویربرداری

یافته‌های این پژوهش را می‌توان به‌صورت منسجم در چارچوب هندسهٔ تصویربرداری و مدل فیزیکی دوربین تفسیر کرد. در مدل سوراخ‌سوزنی، رابطهٔ میان مختصات تصویر و عمق صحنه به‌صورت مستقیم به پارامترهای کانونی و هندسهٔ پروجکشن وابسته است. از این رو، عمق متریک ذاتاً یک کمیت هندسی است و بازیابی آن بدون اطلاع از پارامترهای دوربین به‌صورت یکتا امکان‌پذیر نیست.

نتایج تجربی این پژوهش نشان دادند که مدل‌هایی که تنها بر اطلاعات تصویری تکیه دارند، در بهترین حالت قادر به یادگیری ساختار نسبی عمق هستند. این رفتار با تحلیل‌های نظری فصل سوم هم‌خوانی دارد، زیرا در غیاب ماتریس مشخصهٔ دوربین، نگاشت تصویر به عمق دارای درجه‌ای از آزادی مقیاس است که شبکهٔ یادگیری عمیق نمی‌تواند آن را صرفاً از دادهٔ تصویری حذف کند. به همین دلیل، پایداری مشاهده‌شده در معیارهای غیرمتریک لزوماً به معنای بازیابی صحیح عمق فیزیکی نیست.

در مقابل، زمانی که اطلاعات هندسی دوربین به‌صورت صریح در فرآیند تخمین عمق لحاظ می‌شود، مدل یادگیری با محدودیت‌های فیزیکی سیستم تصویربرداری هم‌راستا می‌گردد. نتایج این پژوهش نشان می‌دهد که ادغام صریح ماتریس مشخصهٔ دوربین در عمل نقش یک قید هندسی را ایفا می‌کند که فضای پاسخ‌های ممکن شبکه را محدود می‌سازد. این محدودسازی باعث می‌شود که مدل به‌جای اتکا به هم‌بستگی‌های آماری ناپایدار، وابستگی‌های هندسی معنادار میان تصویر و عمق را بیاموزد.

تحلیل رفتار معیارهای ارزیابی نیز از این تفسیر هندسی پشتیبانی می‌کند. معیارهای نرمال‌شده، به‌دلیل حذف ضمنی اثر مقیاس، نمی‌توانند تغییرات ناشی از پارامترهای دوربین را به‌خوبی منعکس کنند، در حالی که معیارهای متریک مستقیماً به روابط پروجکشن وابسته‌اند و نسبت به خطاهای هندسی حساس هستند. بهبود پایداری در

این معیارها نشان‌دهندهٔ آن است که مدل پیشنهادی (دارای K) توانسته است وابستگی خروجی عمق به هندسهٔ تصویربرداری را به‌صورت سازگارتر مدل‌سازی کند.

از منظر هندسی، یافته‌های این پژوهش تأیید می‌کنند که یادگیری عمیق، بدون لحاظ صریح مدل تصویربرداری، نمی‌تواند جایگزین کامل قیود فیزیکی شود. در عوض، ترکیب آگاهانهٔ یادگیری داده‌محور با مدل‌های کلاسیک هندسی می‌تواند به تخمین‌هایی منجر شود که هم از نظر بصری معنادارند و هم از نظر فیزیکی پایدار. این تفسیر نشان می‌دهد که دستاورد اصلی پژوهش حاضر نه صرفاً بهبود یک مدل خاص، بلکه ارائهٔ شواهد تجربی برای اهمیت ادغام هندسهٔ تصویربرداری در طراحی مدل‌های تخمین عمق تک‌دوربینه است.

۴.۵ نوآوری‌های تحقیق

بر اساس یافته‌های نظری و تجربی ارائه‌شده در این پژوهش، نوآوری‌های تحقیق حاضر را می‌توان در دو سطح تئوریک و عملی دسته‌بندی کرد. این نوآوری‌ها نه بر مبنای ادعاهای کلی، بلکه بر اساس نتایج عملی استخراج‌شده و تحلیل‌های انجام‌شده در فصل‌های سوم و چهارم تشریح می‌شوند.

۱.۴.۵ نوآوری تئوریک

نوآوری تئوریک اصلی این پژوهش در بازتعریف مسئلهٔ تخمین عمق تک‌دوربینه در چارچوب هندسهٔ تصویربرداری نهفته است. در بسیاری از کارهای پیشین، تخمین عمق به‌عنوان یک مسئلهٔ صرفاً داده‌محور مدل‌سازی شده و نقش پارامترهای هندسی دوربین یا نادیده گرفته شده یا به‌صورت ضمنی در داده‌ها فرض شده است.

در این پژوهش نشان داده شد که ابهام مقیاس در تخمین عمق تک‌دوربینه یک محدودیت ذاتی هندسی است و نه صرفاً یک ضعف در معماری شبکه یا فرآیند آموزش. بر این اساس، نوآوری تئوریک تحقیق حاضر در این است که بازیابی عمق متریک پایدار تنها در صورتی امکان‌پذیر است که اطلاعات هندسی دوربین به‌صورت صریح وارد فرآیند یادگیری شود.

افزون بر این، پژوهش حاضر با تفکیک روشن میان معیارهای متریک و غیرمتریک و تحلیل رفتار متفاوت آن‌ها نشان می‌دهد که ارزیابی‌های متداول مبتنی بر معیارهای نرمال‌شده می‌توانند تصویری ناقص یا حتی گمراه‌کننده از عملکرد واقعی مدل‌ها ارائه دهند. این دیدگاه، چارچوبی مفهومی برای بازنگری در نحوهٔ ارزیابی مدل‌های تخمین عمق در ادبیات موجود فراهم می‌کند.

۲.۴.۵ نوآوری عملی

در سطح عملی، نوآوری این پژوهش در ارائهٔ یک راهکار ساده، قابل پیاده‌سازی و سازگار با معماری‌های موجود برای بهبود تخمین عمق متریک است. برخلاف بسیاری از روش‌ها که نیازمند تغییرات گسترده در معماری یا استفاده از داده‌های اضافی هستند، روش پیشنهادی با حفظ هستهٔ معماری پایه و افزودن صریح اطلاعات ماتریس مشخصهٔ دوربین در قالب یک توکن دروازه‌بانی‌شده به انکودر شبکه، توانسته است پایداری متریک خروجی عمق را افزایش دهد.

یافته‌های عملی این پژوهش نشان می‌دهد که حتی بدون تغییر انکودر و با ثابت نگه‌داشتن وزن‌های استخراج ویژگی، می‌توان با بهره‌گیری از اطلاعات هندسی مناسب، به بهبود معنادار در معیارهای متریک دست یافت. این نتیجه، توصیه‌ای عملی برای پژوهشگران و کاربردهای صنعتی است که نشان می‌دهد نادیده گرفتن اطلاعات دوربین می‌تواند منجر به نتایج ناپایدار در کاربردهایی شود که به مقیاس فیزیکی دقیق نیاز دارند.

در مجموع، نوآوری عملی این پژوهش در ارائهٔ شواهد تجربی است که نشان می‌دهد ادغام آگاهانهٔ هندسهٔ دوربین در مدل‌های یادگیری عمیق می‌تواند بدون افزایش قابل توجه پیچیدگی سیستم، کیفیت و پایداری تخمین عمق متریک را بهبود بخشد.

۳.۴.۵ دستاوردهای روش‌شناختی و تشخیصی

در کنار نوآوری‌های فوق، این پژوهش مجموعه‌ای از دستاوردهای روش‌شناختی دارد که اعتبار نتایج را تضمین کرده و برای پژوهش‌های مشابه قابل استفاده‌اند:

- **شناسایی و رفع نشت داده در تقسیم‌بندی NYUv2:** نشان داده شد که تقسیم تصادفی متداول بر روی مجموعهٔ برچسب‌دار، بیش از ۸۰ درصد تصاویر آزمون استاندارد را وارد آموزش می‌کند؛ اصلاح این نشت، پیش‌شرط هر مقایسهٔ معتبر با ادبیات موضوع است.

- **تشخیص و تبیین یک شکست خاموش در فاین‌تیون مدل‌های نسبی:** ناهم‌خوانی فضای هدف آموزش (عمق مستقیم) با فضای خروجی مدل ازپیش‌آموزش‌دیده (عمق معکوس)، در تعامل با جزئیات تابع هزینهٔ تغییرناپذیر نسبت به مقیاس، به فروریزش برگشت‌ناپذیر سر شبکه می‌انجامد در حالی که منحنی هزینه طبیعی می‌نماید. ریشه‌یابی و رفع این پدیده، آموزش موفق مدل‌های بزرگ را ممکن ساخت.

- **افزونه‌سازی هندسی آگاه از دوربین:** راهبردی که با برش‌های مقیاس‌متغیر و به‌روزرسانی دقیق ماتریس K، از یک مجموعه‌دادهٔ تک‌دوربینه، داده‌هایی با میدان دید متنوع و برچسب هندسی صحیح می‌سازد.

این راهبرد به‌تنهایی بهترین نتایج پروتکل نسبی این پژوهش را رقم زد و بستر آزمون علّی سهم پارامترهای دوربین را فراهم کرد.

- **تحلیل وابستگی نتیجه‌گیری به پروتکل ارزیابی:** نشان داده شد که پروتکل ارزیابی نسبی، به‌صورت ساختاری نسبت به سهم اصلی پارامترهای دوربین (مقیاس مطلق) نابیناست و سنجش منصفانهٔ این سهم، مستلزم ارزیابی متریک بدون تراز است. این یافته برای طراحی ارزیابی در پژوهش‌های آتی حوزه اهمیت دارد.

- **افزونه‌سازی لرزش کانونی:** تبدیل دقیق $(f, Z) \mapsto (\alpha f, \alpha Z)$ برگرفته از معادلهٔ روزنه‌ای که بدون هیچ بازنمونه‌برداری، بازیابی مقیاس را برای مدل فاقد K به مسئله‌ای بدساخت تبدیل می‌کند و اتکای مدل به پارامترهای دوربین را از یک گزینه به یک ضرورت آموزشی ارتقا می‌دهد (زیربخش ۸.۷.۳).

- **آزمون تشخیصی جاروی K:** سنجه‌ای سازوکاری (شیب لگاریتمی پاسخ عمق به مقیاس‌دهی مصنوعی فاصلهٔ کانونی) که مستقیماً اندازه می‌گیرد مدل چقدر K را «می‌خواند»، و نشان داد که معیارهای تجمیعی و حتی مقدار پارامتر دروازهٔ کدگذار دوربین می‌توانند دربارهٔ میزان واقعی مصرف K گمراه‌کننده باشند (زیربخش ۱.۶.۴).

- **زوج‌سازی کامل آزمایش‌ها با قفل بذر تصادفی:** یکسان‌سازی ترتیب داده، وارونه‌سازی‌ها و نمونه‌های افزونه‌سازی میان بازوهای مدل پیشنهادی (دارای K) و مدل کنترل (فاقد K)، به‌گونه‌ای که تفاوت نتایج تنها به متغیر مورد مطالعه (ورودی K) قابل انتساب باشد؛ صحت آن با برابری بیت‌به‌بیت هزینهٔ گام صفر دو بازو راستی‌آزمایی شد.

جدول ۱.۵ این نوآوری‌ها را در سه سطح تئوریک، عملی و روش‌شناختی جمع‌بندی می‌کند و برای هر مورد، شاهد پشتیبان و قابلیت استفادهٔ مجدد آن در پژوهش‌های دیگر را مشخص می‌سازد.

۵.۵ محدودیت‌ها و چالش‌های پژوهش

با وجود نتایج مثبت و معنادار حاصل از این پژوهش، لازم است محدودیت‌ها و چالش‌هایی که در مسیر انجام تحقیق وجود داشته‌اند به‌صورت شفاف بیان شوند. ذکر این موارد نه‌تنها به درک دقیق‌تر دامنهٔ اعتبار یافته‌ها کمک می‌کند، بلکه مسیر توسعه و بهبود پژوهش‌های آتی در این حوزه را نیز روشن‌تر می‌سازد.

جدول ۱.۵: جمع‌بندی نوآوری‌ها و دستاوردهای پژوهش در سه سطح تئوریک، عملی و روش‌شناختی، به‌همراه شاهد پشتیبان هر مورد.

سطح	دستاورد	شرح	شاهد پشتیبان
تئوریک	بازتعریف ابهام مقیاس به‌عنوان محدودیت هندسی	ابهام مقیاس یک ویژگی ذاتی نگاشت تصویربرداری است، نه ضعف معماری یا کمبود داده؛ بنابراین با بزرگ‌ترکردن شبکه رفع نمی‌شود	شکل ۱.۱ و آزمون حذف K (۲.۶.۴)
تئوریک	نابینایی ساختاری پروتکل نسبی	تراز مقیاس به‌ازای هر تصویر، دقیقاً همان کمیتی را حذف می‌کند که K فراهم می‌آورد؛ پس ارزیابی نسبی نمی‌تواند سهم K را بسنجد	بخش ۵.۵.۴
عملی	شرطی‌سازی توکنی با دروازهٔ صفرمقدار	افزودن K بدون تغییر هستهٔ معماری و بدون آسیب به وزن‌های ازپیش‌آموخته، حتی با انکودر کاملاً فریزشده	شکل ۳.۳
روش‌شناختی	افزونه‌سازی هندسی آگاه از دوربین	ساخت داده‌های با میدان دید متنوع و برچسب هندسی صحیح از یک مجموعه‌دادهٔ تک‌دوربینه	شکل ۵.۳؛ بهترین نتایج پروتکل نسبی
روش‌شناختی	افزونه‌سازی لرزش کانونی	تبدیل دقیق $(f, Z) \mapsto (\alpha f, \alpha Z)$ که بازیابی مقیاس را برای مدل فاقد K بدساخت می‌کند، بدون هیچ بازنمونه‌برداری یا نشت اطلاعاتی	شکل ۶.۳ و جدول ۱۵.۴
روش‌شناختی	آزمون تشخیصی جاروی K	سنجه‌ای مستقیم برای اینکه مدل چقدر واقعاً K را «می‌خواند»؛ نشان داد معیارهای تجمیعی و حتی مقدار دروازهٔ کدگذار می‌توانند گمراه‌کننده باشند	شکل ۹.۴
روش‌شناختی	زوج‌سازی کامل با قفل بذر تصادفی	یکسان‌سازی ترتیب داده، وارونه‌سازی‌ها و نمونه‌های افزونه‌سازی میان دو بازو، تا تفاوت نتایج تنها به ورودی K قابل انتساب باشد	برابری بیت‌به‌بیت هزینهٔ گام صفر دو بازو
روش‌شناختی	مستندسازی سه شکست خاموش	نشت تقسیم‌بندی، ناهم‌خوانی فضای هدف، و ناهم‌خوانی معماری سر شبکه؛ هر سه بدون هیچ پیام خطایی رخ می‌دادند	جدول ۶.۳

نخستین محدودیت این پژوهش به وابستگی روش پیشنهادی به صحت ماتریس مشخصهٔ دوربین بازمی‌گردد. در این تحقیق فرض شده است که پارامترهای درونی دوربین به‌صورت دقیق و بدون خطا در دسترس هستند. آزمون حذف و تحریف K (زیربخش ۲.۶.۴) ابعاد این وابستگی را کمی کرده است: حذف کامل K حدود ۵۹ درصد و تحریف ۳۰ تا ۴۰ درصدیِ فاصلهٔ کانونی، چند برابر خطای پایه به مدل تحمیل می‌کند. این رفتار، روی دیگر سکهٔ اتکای علّی مدل به K است: در عمل، به‌ویژه در کاربردهای واقعی که پارامترها ممکن است ناقص، تقریبی یا همراه با نویز باشند، حساسیت مدل به خطای کالیبراسیون باید در طراحی سامانه لحاظ شود؛ مطالعهٔ نظام‌مند رفتار مدل تحت عدم قطعیت‌های واقع‌گرایانهٔ کالیبراسیون (نه تحریف‌های عمدی بزرگ) موضوعی برای پژوهش‌های آتی است.

محدودیت دیگر به کیفیت و ماهیت داده‌های عمق مرجع در دیتاست‌های مختلف مربوط است. همان‌گونه که در فصل چهارم نشان داده شد، دیتاست‌ها از نظر نحوهٔ تولید عمق مرجع، دامنهٔ عمق و یکنواختی داده‌ها تفاوت‌های اساسی دارند. این ناهمگنی می‌تواند بر مقادیر مطلق خطاهای متریک اثرگذار باشد و تفسیر نتایج را تحت تأثیر

قرار دهد. هرچند در این پژوهش تلاش شده است این مسئله از طریق تحلیل‌های آماری و بررسی‌های موردی کنترل شود، اما محدودیت‌های ذاتی داده‌های مرجع به‌طور کامل قابل حذف نیستند.

از منظر معماری، روش پیشنهادی بر پایهٔ یک مدل خاص از خانوادهٔ Depth Anything V2 پیاده‌سازی و ارزیابی شده است. اگرچه چارچوب ارائه‌شده به‌صورت مفهومی مستقل از معماری خاص است، اما تعمیم تجربی نتایج به سایر معماری‌ها یا انکودرهای متفاوت در این پژوهش به‌صورت مستقیم بررسی نشده است. از این رو، نتایج ارائه‌شده بیشتر در چارچوب معماری مورد استفاده قابل تفسیر هستند.

در نهایت، این پژوهش تمرکز خود را بر تخمین عمق تک‌دوربینهٔ ایستا و تصاویر منفرد قرار داده است. اطلاعات زمانی یا چندنمایی که می‌توانند قیود هندسی مکملی فراهم کنند، در چارچوب این تحقیق مورد استفاده قرار نگرفته‌اند. این انتخاب آگاهانه، برای تمرکز بر نقش ماتریس مشخصهٔ دوربین در ساده‌ترین سناریوی ممکن انجام شده است، اما در عین حال دامنهٔ کاربرد نتایج را به این سناریو محدود می‌کند.

سه محدودیت دیگر نیز که از مطالعهٔ کنترل‌شدهٔ نهایی برمی‌خیزند باید به‌صراحت ذکر شوند. نخست، تنوع پارامترهای دوربین در آزمایش‌های نهایی از طریق افزونه‌سازی هندسی (برش‌های مقیاس‌متغیر با K اصلاح‌شده) شبیه‌سازی شده است، نه با داده‌های واقعی چنددوربینه؛ اگرچه این شبیه‌سازی از منظر هندسی دقیق است، اثرات ثانویهٔ دوربین‌های واقعی (نظیر اعوجاج لنز، عمق میدان و نویز حسگر) را پوشش نمی‌دهد. دوم، در دو دور متریکِ مبتنی بر زوم‌ـبرش، برتری مدل پیشنهادی (دارای K) در RMSE تثبیت شد، اما AbsRel دور دوم در آستانهٔ معناداری بود (ویلکاکسون $p = ۰/۰۴۰$ و آزمون t زوجی $p = ۰/۰۸۹$). این محدودیت مربوط به آن دو دور است؛ در آزمایش لرزش کانونی تمام معیارها با $p \leq ۱۰^{-۲۴}$ معنادار شدند. بنابراین قدرت شواهد به میزان ابهام هندسی ایجادشده در داده وابسته است. سوم، مقایسه‌های کنترل‌شدهٔ نهایی بر یک محک درون‌دامنه‌ای (NYUv2) متمرکز بوده‌اند؛ تعمیم سهم پارامترهای دوربین به فاین‌تیون بر محک‌های دیگر (به‌ویژه صحنه‌های بیرونی با دامنهٔ عمق بزرگ) نیازمند آزمایش‌های تکمیلی است. در صورت دسترسی به دادهٔ آموزشی واقعی KITTI، تکرار مطالعهٔ کنترل‌شدهٔ متریک بر روی آن می‌تواند تعمیم‌پذیری یافته‌ها را در محیط بیرونی و دامنهٔ عمق بزرگ مستند کند.

چهارم، هر بازوی هر زوج آموزشی تنها یک بار آموزش داده شده است. قفل بذر، ترتیب داده و افزونه‌سازی را میان دو بازو یکسان می‌کند و آزمون‌های زوجی عدم قطعیت در سطح تصویر را می‌سنجند، اما جایگزین تکرار کامل آموزش با چند بذر مستقل نیست. ازاین‌رو، استنباط علّی درون هر زوج قوی است، ولی برآورد واریانس ناشی از تصادفی‌بودن کل فرآیند آموزش همچنان به آزمایش تکمیلی نیاز دارد.

با در نظر گرفتن این موارد، می‌توان گفت که محدودیت‌های مطرح‌شده نتایج اصلی پژوهش را نقض نمی‌کنند، بلکه چارچوب اعتبار آن‌ها را مشخص می‌سازند و زمینه‌ای مناسب برای توسعه و تکمیل پژوهش در مطالعات آینده فراهم می‌کنند.

جدول ۲.۵ این محدودیت‌ها را در کنار اثر هر یک بر دامنهٔ اعتبار یافته‌ها و مسیر پژوهشی متناظر برای رفع آن قرار می‌دهد.

جدول ۲.۵: نگاشت محدودیت‌های شناسایی‌شدهٔ پژوهش به اثر هر یک بر دامنهٔ اعتبار یافته‌ها و مسیر پژوهشی متناظر برای رفع آن.

محدودیت	اثر بر دامنهٔ اعتبار یافته‌ها	مسیر پژوهشی پیشنهادی
وابستگی به صحت ماتریس K	همان اتکای علّی که مزیت مدل است، آن را به خطای کالیبراسیون حساس می‌کند	مطالعهٔ رفتار مدل تحت عدم قطعیت‌های واقع‌گرایانهٔ کالیبراسیون؛ طراحی سازوکار مقاوم به نویز K
شبیه‌سازی تنوع دوربین با افزونه‌سازی	اثرات ثانویهٔ دوربین‌های واقعی مانند اعوجاج لنز، عمق میدان و نویز حسگر پوشش داده نشده است	آموزش بر داده‌های واقعاً چنددوربینه؛ الگوی پاسخ به دوز پیش‌بینی می‌کند سهم K در این حالت بزرگ‌تر شود
تمرکز بر یک محک درون‌دامنه‌ای	تعمیم یافته‌ها به صحنه‌های بیرونی با دامنهٔ عمق بزرگ آزموده نشده است	تکرار مطالعهٔ کنترل‌شدهٔ متریک روی KITTI یا محک‌های بیرونی مشابه
تک‌اجرا بودن هر بازو	استنباط علّی درون هر زوج قوی است، اما واریانس ناشی از تصادفی‌بودن کل فرآیند آموزش برآورد نشده است	تکرار هر زوج با چند بذر مستقل و گزارش میانگین و انحراف معیار میان اجراها
تک‌معماری بودن ارزیابی	نتایج در چارچوب خانوادهٔ Depth Anything V2 تفسیرپذیرند	پیاده‌سازی همان چارچوب روی انکودرها و دیکودرهای متفاوت
عدم پیمایش محل تزریق توکن	تزریق تنها در نخستین لایهٔ انکودر آزموده شده است	پیمایش نظام‌مند عمق تزریق و بررسی تزریق چندلایه
نبود پروتکل کیفی قفل‌شده	نتیجه‌گیری به شواهد کمی و آزمون‌های علّی محدود مانده است	قفل‌کردن شناسهٔ تصاویر پیش از مشاهده و ذخیرهٔ خروجی هر دو بازو با نگاشت رنگ و بازهٔ عمق یکسان

۶.۵ پیشنهادهایی برای تحقیقات آینده

یافته‌ها و محدودیت‌های مطرح‌شده در این پژوهش، زمینه‌های روشنی را برای ادامه و توسعهٔ تحقیقات آینده در حوزهٔ تخمین عمق تک‌دوربینه فراهم می‌کنند. پیشنهادهای ارائه‌شده در این بخش به‌طور مستقیم بر چارچوب مفهومی و عملی تحقیق حاضر استوار هستند و از طرح موضوعات کلی و خارج از دامنهٔ پژوهش اجتناب شده است.

یکی از مسیرهای طبیعی برای توسعهٔ این پژوهش، بررسی رفتار مدل پیشنهادی (دارای K) در شرایطی است که ماتریس مشخصهٔ دوربین به‌صورت دقیق در دسترس نیست. در بسیاری از کاربردهای واقعی، پارامترهای

درونی دوربین ممکن است همراه با نویز یا به‌صورت تخمینی ارائه شوند. مطالعهٔ پایداری مدل پیشنهادی (دارای K) در برابر خطاهای موجود در ماتریس K و یا طراحی سازوکارهایی برای مقابله با این عدم قطعیت، می‌تواند گام مهمی در جهت کاربردپذیری عملی‌تر روش ارائه‌شده باشد.

مسیر پژوهشی دیگر، یادگیری هم‌زمان عمق و پارامترهای دوربین است. در چارچوب تحقیق حاضر، ماتریس مشخصهٔ دوربین به‌عنوان ورودی معلوم در نظر گرفته شده است. بررسی امکان تخمین یا اصلاح پارامترهای درونی دوربین به‌صورت هم‌زمان با تخمین عمق، می‌تواند منجر به مدل‌هایی شود که وابستگی کمتری به کالیبراسیون اولیه دارند و در سناریوهای متنوع‌تری قابل استفاده هستند.

مسیر مهم دیگری که مستقیماً از یافته‌های مطالعهٔ کنترل‌شده برمی‌خیزد، آموزش بر روی داده‌های واقعاً چنددوربینه است. الگوی پاسخ به دوز مشاهده‌شده نشان می‌دهد که سهم پارامترهای دوربین با میزان تنوع هندسی داده رشد می‌کند؛ بنابراین انتظار می‌رود آموزش بر ترکیبی از مجموعه‌داده‌هایی با دوربین‌ها و میدان‌های دید متفاوت (به‌جای شبیه‌سازی تنوع از طریق افزونه‌سازی) این سهم را بزرگ‌تر و پایدارتر آشکار کند. در همین راستا، مطالعهٔ نظام‌مند محل تزریق توکن دوربین در عمق‌های مختلف انکودر (از طریق پارامتر `--cam-token-inject-layer`) و بررسی سازوکارهای تزریق چندلایه، می‌تواند اثرگذاری اطلاعات هندسی را افزایش دهد.

از منظر استحکام آماری، تکرار هر زوج مدل پیشنهادی (دارای K)/مدل کنترل (فاقد K) با چند بذر مستقل و گزارش میانگین و انحراف معیار میان اجراها ضروری است. این تکرار باید در کنار آزمون‌های زوجی سطح تصویر انجام شود تا دو منبع عدم قطعیت — تفاوت دشواری تصاویر و تصادفی‌بودن آموزش — به‌صورت جداگانه اندازه‌گیری شوند.

همچنین، تعمیم چارچوب پیشنهادی به سناریوهای چندتصویری یا ویدئویی می‌تواند زمینهٔ پژوهشی ارزشمندی باشد. اطلاعات زمانی یا چندنمایی، قیود هندسی مکملی نسبت به اطلاعات تک‌تصویری فراهم می‌کنند و ترکیب آن‌ها با ادغام صریح ماتریس مشخصهٔ دوربین می‌تواند به تخمین‌های پایدارتر و دقیق‌تر از عمق متریک منجر شود.

از منظر معماری، بررسی پیاده‌سازی روش پیشنهادی بر روی انکودرها و دیکودرهای متفاوت و تحلیل میزان وابستگی نتایج به انتخاب معماری پایه می‌تواند به تعمیم‌پذیری بهتر چارچوب ارائه‌شده کمک کند. چنین بررسی‌هایی امکان ارزیابی این موضوع را فراهم می‌کنند که آیا مزایای مشاهده‌شده یک ویژگی عمومی رویکرد هندسی پیشنهادی هستند یا به معماری خاصی وابسته‌اند.

در مجموع، پیشنهادهای مطرح‌شده در این بخش نشان می‌دهند که پژوهش حاضر می‌تواند به‌عنوان نقطهٔ آغاز برای مجموعه‌ای از مطالعات عمیق‌تر در زمینهٔ ترکیب هندسهٔ تصویربرداری و یادگیری عمیق در تخمین عمق تک‌دوربینه مورد استفاده قرار گیرد.

۷.۵ جمع‌بندی نهایی فصل

در این فصل، جمع‌بندی نهایی پژوهش حاضر با تمرکز بر برداشت‌های حاصل از یافته‌ها ارائه شد. بر اساس نتایج نظری و تجربی به‌دست‌آمده، نشان داده شد که تخمین عمق متریک پایدار در چارچوب تک‌دوربینه بدون لحاظ صریح هندسهٔ دوربین به‌صورت ذاتی با محدودیت مواجه است. این برداشت، ریشه در اصول هندسهٔ تصویربرداری دارد و با نتایج عملی گزارش‌شده در فصل‌های پیشین هم‌راستا است.

یافته‌های این پژوهش نشان می‌دهد که ادغام آگاهانهٔ ماتریس مشخصهٔ دوربین در فرآیند تخمین عمق می‌تواند نقش تعیین‌کننده‌ای در کاهش ابهام مقیاس و افزایش پایداری متریک خروجی ایفا کند. این بهبودها نه‌تنها در معیارهای ارزیابی، بلکه در کیفیت بصری نقشه‌های عمق و رفتار مدل در سناریوهای متنوع تصویربرداری قابل مشاهده بودند.

در مجموع، پژوهش حاضر نشان می‌دهد که ترکیب یادگیری عمیق با قیود هندسی کلاسیک، مسیر مؤثری برای ارتقای تخمین عمق تک‌دوربینه است. نتایج ارائه‌شده می‌توانند به‌عنوان مبنایی برای توسعهٔ مدل‌های آتی و بازنگری در شیوهٔ ارزیابی روش‌های تخمین عمق مورد استفاده قرار گیرند. بدین ترتیب، این پایان‌نامه با ارائهٔ یک چارچوب نظری و عملی منسجم، پاسخی روشن به مسئلهٔ مطرح‌شده در ابتدای تحقیق ارائه می‌دهد و زمینه را برای پژوهش‌های تکمیلی در این حوزه فراهم می‌سازد.

کتاب‌نامه

[1] Saxena, Ashutosh, Sun, Min, and Ng, Andrew Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009.

[2] Eigen, David, Puhrsch, Christian, and Fergus, Rob. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 2014.

[3] Hartley, Richard and Zisserman, Andrew. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[4] Koch, Reinhard. Dynamic 3-d scene analysis through synthesis feedback control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[5] Eigen, David and Fergus, Rob. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *IEEE International Conference on Computer Vision*, 2015.

[6] Ranftl, René, Bochkovskiy, Alexey, and Koltun, Vladlen. Vision transformers for dense prediction. *IEEE International Conference on Computer Vision*, 2021.

[7] Zhou, Tinghui, Brown, Matthew, Snavely, Noah, and Lowe, David. Unsupervised learning of depth and ego-motion from video. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[8] Guizilini, Vitor et al. Semantically-guided representation learning for self-supervised monocular depth. *IEEE International Conference on Computer Vision*, 2020.

[9] Godard, Clément, Mac Aodha, Oisin, and Brostow, Gabriel. Unsupervised monocular depth estimation with left-right consistency. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[10] Hoiem, Derek, Efros, Alexei, and Hebert, Martial. Geometric context from a single image. *IEEE International Conference on Computer Vision*, 2005.

[11] Ranftl, Rene et al. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint*, 2019.

[12] Saxena, Shubham et al. Zero-shot metric depth with a field-of-view conditioned diffusion model. *arXiv preprint arXiv:2312.13252*, 2023.

[13] Yang, Lihe et al. Depth anything v2. *arXiv preprint*, 2024.

[14] Godard, Clément, Aodha, Oisin Mac, Firman, Michael, and Brostow, Gabriel J. Digging into self-supervised monocular depth estimation. in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[15] Guizilini, Vitor et al. 3d packing for self-supervised monocular depth estimation. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] Chen, Rui et al. Camera-aware multi-view stereo via geometry-aware cost volume. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Dosovitskiy, Alexey et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[18] Casser, Vincent, Pirk, Soren, Mahjourian, Reza, and Angelova, Anelia. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[19] Sucar, Edgar, Liu, Shichen, Ortiz, Joseph, and Davison, Andrew J. Implicit representation of geometry for learning depth. in *Proceedings of the Conference on Robot Learning (CoRL)*, 2020.

[20] Kendall, Alex and Cipolla, Roberto. Geometry-aware learning of maps for camera localization. in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[21] Guizilini, Vitor, Ambrus, Rares, Pillai, Sudeep, and Gaidon, Adrien. Semantically-guided representation learning for self-supervised monocular depth. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Perez, Ethan, Strub, Florian, de Vries, Harm, Dumoulin, Vincent, and Courville, Aaron. Film: Visual reasoning with a general conditioning layer. in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[23] Li, Zhengqi et al. Neural scene flow fields for space-time view synthesis of dynamic scenes. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[24] Gaidon, Adrien, Wang, Qiao, Cabon, Yohann, and Vig, Eleonora. Virtual worlds as proxy for multi-object tracking analysis. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Silberman, Nathan, Hoiem, Derek, Kohli, Pushmeet, and Fergus, Rob. Indoor segmentation and support inference from rgbd images. *European Conference on Computer Vision (ECCV)*, 2012.

[26] Vasiljevic, Igor and Koltun, Vladlen. Diode: A dense indoor and outdoor depth dataset. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] Oquab, Maxime et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[28] Loshchilov, Ilya and Hutter, Frank. Decoupled weight decay regularization. in *International Conference on Learning Representations (ICLR)*, 2019.

*واژه‌نامهٔ فارسی به انگلیسی

*واژه‌نامهٔ انگلیسی به فارسی

Abstract

Monocular depth estimation is intrinsically ambiguous because a single image does not uniquely determine the metric scale of a scene. This thesis investigates whether supplying camera intrinsics explicitly can reduce that ambiguity in a foundation depth model. A camera-aware extension of Depth Anything V2 is developed in which a normalized nine-dimensional camera descriptor is encoded as a gated token and injected into the DINOv2 encoder. The experimental methodology was refined through a sequence of controlled investigations that exposed and corrected data leakage in the NYUv2 split, a silent disparity–depth target mismatch that collapsed ViT-L training, geometrically inconsistent crop augmentation, and a silent metric-head mismatch during evaluation.

All final comparisons use the leak-free 654-image NYUv2 Eigen test set and matched camera-aware and camera-blind training arms. Under relative evaluation, the proposed ViT-S model improves over both the released model and the matched control, while the ViT-L camera contribution is not significant because per-image scale-and-shift alignment removes the principal information supplied by the focal length. Under metric evaluation without alignment, the camera-aware ViT-L model consistently improves over its matched control, and the effect grows as field-of-view diversity increases. A geometrically exact focal-jitter augmentation is then introduced to make scale unrecoverable from image pixels alone. With this augmentation, the camera-aware model obtains AbsRel 0.0849 versus 0.1030 for the control and improves every reported metric with Wilcoxon significance of at least $p \leq 10^{-24}$.

Mechanistic and causal tests confirm that the improvement is genuinely attributable to the intrinsic matrix. The log–log slope of predicted depth versus injected focal scale rises from 0.06 before focal jitter to 0.93–0.97 afterward, close to the pinhole-model ideal of one. Withholding the intrinsic matrix increases AbsRel by 59 percent, while deliberately corrupting focal length produces errors of the magnitude predicted by projective geometry. These results show that camera intrinsics can provide a statistically significant and causally necessary scale signal, while also revealing sensitivity to calibration accuracy and the need for future evaluation on genuinely multi-camera and outdoor data.

Keywords Monocular Depth Estimation, Unsupervised Deep Learning, Metric Depth, Camera Intrinsics, Imaging Geometry, Depth Anything V2

K. N. Toosi University of Technology
Faculty of Computer Engineering- Artificial Intelligence Group

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of Master of Science (M.Sc.)
in Computer Engineering

Unsupervised Depth Estimation Using Internal Calibration Parameters as Input

By:

Mohammad Amin Ghasemzadeh

Supervisor:

Dr. Behrouz Nasihatkon

Winter 2026