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

زمستان ۱۴۰۴

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

زمستان ۱۴۰۴

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 زمستان ۱۴۰۴مورد 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

۴. نمایندهٔ تحصیلات تکمیلی دانشکده: دکتر نماینده. . . . . امضا

دانشگاه صنعتی خواجه نصیرالدین طوسی

اظهارنامه دانشجو

اینجانب محمدامین قاسم‌زاده به شماره دانشجویی ۴۰۲۰۹۶۴۴دانشجوی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر دانشکده ... 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

زمستان ۱۴۰۴

چکیده

تخمین عمق تک‌دوربینه یکی از مسائل بنیادین در بینایی ماشین است که کاربردهای گسترده‌ای در رباتیک، رانندگی خودکار و بازسازی سه‌بعدی دارد. با وجود پیشرفت‌های اخیر مبتنی بر یادگیری عمیق، بسیاری از روش‌های موجود همچنان با ناپایداری مقیاس متریک و حساسیت بالا نسبت به مشخصات دوربین مواجه‌اند. به‌ویژه، اغلب مدل‌های یادگیری‌محور بدون در نظر گرفتن صریح هندسهٔ تصویربرداری، عمق را به‌صورت نسبی یا وابسته به دیتاست تخمین می‌زنند.

در این پژوهش، مسئلهٔ تخمین عمق متریک پایدار در چارچوب تک‌دوربینه بررسی شده و نقش پارامترهای درونی دوربین در پایداری خروجی‌های عمق مورد تحلیل قرار گرفته است. بدین منظور، نسخه‌ای اصلاح‌شده از مدل ۲V Anything Depth ارائه می‌شود که در آن ماتریس درونی دوربین به‌صورت صریح و ساخت‌یافته در فرآیند تخمین عمق ادغام شده است. هدف اصلی این رویکرد، کاهش حساسیت مدل نسبت به تغییرات کالیبراسیون دوربین و افزایش سازگاری متریک خروجی‌ها در شرایط متنوع تصویربرداری است.

ارزیابی‌های تجربی گسترده‌ای بر روی دیتاست‌های VKITTI، ۲NYUv، DIODE Stereo Driving و CityScapes انجام شده است. نتایج کمی نشان می‌دهد که مدل پیشنهادی در مقایسه با مدل پایه، بهبود معناداری در معیارهای متریک از جمله RMSE و SiLog ارائه می‌دهد. آزمون‌های آماری شامل آزمون t و بازه‌های اطمینان bootstrap معناداری آماری این بهبودها را تأیید می‌کنند. همچنین تحلیل حساسیت نشان می‌دهد که معیارهایی نظیر AbsRel توان آشکارسازی ناپایداری‌های مقیاس را ندارند، در حالی که RMSE واکنش مستقیم‌تری به تغییرات هندسی دوربین نشان می‌دهد.

در مجموع، نتایج این پژوهش بیانگر آن است که ادغام صریح هندسهٔ تصویربرداری با مدل‌های یادگیری عمیق می‌تواند چارچوبی مؤثر و قابل اتکا برای دستیابی به تخمین عمق متریک پایدار در تنظیمات تک‌دوربینه فراهم کند.

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 د

فهرست برنامه‌ها ر

فهرست اختصارات ز

فصل ۱: مقدمه ۱

۱.۱ بیان مسئله و اهمیت آن ۱

۲.۱ چالش‌های عمق متریک در مدل‌های یادگیری عمیق ۲

۳.۱ خلأ پژوهشی ۴

۴.۱ اهداف و سوالات تحقیق ۵

۵.۱ ساختار پایان‌نامه ۵

فصل ۲: مرور ادبیات ۷

۱.۲ مفاهیم و تعاریف ۷

۱.۱.۲ تخمین عمق تک‌دوربینه ۷

۲.۱.۲ عمق متریک در مقابل عمق غیرمتریک ۸

۳.۱.۲ ابهام مقیاس ۸

۴.۱.۲ پارامترهای درونی دوربین ۸
۵.۱.۲ ترنسفورمر بینایی ۹
۶.۱.۲ تقطیر دانش ۹
۷.۱.۲ یادگیری خودنظارتی ۱۰
۲.۲ مرور ادبیات و پیشینه پژوهش ۱۰
۱.۲.۲ رویکردهای کلاسیک و یادگیری مبتنی بر ویژگی‌های دستی ۱۱
۲.۲.۲ تحول در تخمین عمق با یادگیری عمیق نظارت‌شده ۱۱
۳.۲.۲ یادگیری خودنظارتی و بازسازی هندسی دیدگاه ۱۱
۴.۲.۲ مدل‌های مبتنی بر ترنسفورمر و یادگیری در مقیاس بزرگ ۱۲
۵.۲.۲ ادغام دانش هندسی و چالش پایداری متریک ۱۲
۳.۲ شکاف تحقیقاتی ۱۲
۱.۳.۲ فقدان پایداری در تخمین عمق متریک ۱۳
۲.۳.۲ نادیده گرفتن صریح هندسه و پارامترهای درونی دوربین ۱۳
۳.۳.۲ حساسیت بالا نسبت به تغییرات دامنه و داده‌های خارج از توزیع ۱۳
۴.۳.۲ محدودیت معیارهای ارزیابی در تفکیک خطاهای هندسی ۱۳
۴.۲ نوآوری‌های تحقیق ۱۴
۱.۴.۲ ادغام صریح پارامترهای درونی دوربین در معماری ترنسفورمر ۱۴
۲.۴.۲ طراحی دیکودر چندمؤلفه‌ای با قابلیت تفکیک ویژگی‌های بصری و هندسی . . . ۱۴
۳.۴.۲ راهبرد آموزش چندمرحله‌ای مبتنی بر تقطیر دانش متریک ۱۴
۴.۴.۲ تحلیل حساسیت سیستماتیک نسبت به تغییرات هندسه تصویربرداری ۱۵
۵.۲ جمع‌بندی فصل ۱۵

فصل ۳: روش تحقیق ۱۷
۱.۳ مقدمه ۱۷
۲.۳ صورت‌بندی مسئلهٔ تخمین عمق تک‌دوربینه بدون نظارت ۱۸
۳.۳ هندسهٔ دوربین و نقش پارامترهای درونی ۲۰

۱.۳.۳ مدل دوربین سوراخ‌سوزنی ۲۰
۲.۳.۳ ماتریس مشخصهٔ دوربین ۲۱
۳.۳.۳ نقش پارامترهای درونی در تخمین عمق ۲۲
۴.۳ معماری پایه: مدل Depth Anything V2 . . . . . ۲۳
۱.۴.۳ ساختار کلی شبکهٔ encoder-decoder ۲۳
۲.۴.۳ بخش encoder مبتنی بر Transformer Vision ۲۳
۳.۴.۳ بخش decoder و راهبرد بازسازی نگاشت عمق ۲۴
۵.۳ انگیزه‌ی صریح برای لحاظ پارامترهای درونی دوربین ۲۵
۱.۵.۳ شواهد تجربی از تاثیر پارامترهای دوربین در معیارهای متریک ۲۵
۲.۵.۳ عدم کفایت رویکرد ضمنی نسبت به مشخصه‌های دوربین ۲۶
۳.۵.۳ کاربردهای عملی و انتقال‌پذیری به داده‌های جهان واقعی ۲۶
۴.۵.۳ لزوم طراحی مدل با ورود صریح پارامترهای درونی ۲۷
۶.۳ طراحی روش پیشنهادی: ادغام صریح ماتریس مشخصهٔ دوربین ۲۷
۱.۶.۳ اصل طراحی: تفکیک اطلاعات تصویری و هندسی ۲۷
۲.۶.۳ نمایش و پیش‌پردازش ماتریس مشخصهٔ دوربین ۲۸
۳.۶.۳ راهبرد ادغام اطلاعات دوربین با معماری Depth Anything V2 . . . . . ۲۹
۴.۶.۳ مزایای مورد انتظار از ادغام صریح پارامترهای درونی ۳۰
۷.۳ راهبرد آموزش مدل پیشنهادی ۳۰
۱.۷.۳ مجموعه‌داده‌های مورد استفاده ۳۱
۲.۷.۳ مدل پایه و راهبرد فاین‌تیون ۳۱
۳.۷.۳ فعال‌سازی پشتیبانی از پارامترهای دوربین ۳۲
۴.۷.۳ تنظیمات آموزش و بهینه‌سازی ۳۲
۵.۷.۳ استفاده از تقطیر دانش ۳۳
۶.۷.۳ جمع‌بندی فرآیند آموزش ۳۳
۸.۳ توابع هزینه و معیارهای بهینه‌سازی ۳۳
۱.۸.۳ هزینهٔ SiLog برای مدل‌های متریک ۳۵

۲.۸.۳ هزینهٔ تغییرناپذیر نسبت به مقیاس و جابجایی برای مدل‌های نسبی ۳۵
۳.۸.۳ تمایز میان توابع هزینهٔ آموزش و معیارهای ارزیابی ۳۶
۱.۳.۸.۳ توابع هزینه در آموزش ۳۶
۲.۳.۸.۳ معیارهای ارزیابی ۳۷
۴.۸.۳ نقش پارامترهای درونی دوربین در آموزش ۳۷
۵.۸.۳ حساسیت RMSE به پارامترهای دوربین و کاربرد آن در ارزیابی ۳۸
۶.۸.۳ جمع‌بندی ۳۸
۹.۳ جزئیات پیاده‌سازی ۳۹
۱.۹.۳ چارچوب نرم‌افزاری و زیرساخت محاسباتی ۳۹
۲.۹.۳ پیش‌پردازش داده‌ها ۳۹
۳.۹.۳ مدیریت و نرمال‌سازی پارامترهای درونی دوربین ۴۰
۴.۹.۳ تنظیمات آموزش و راهبرد بهینه‌سازی ۴۰
۵.۹.۳ دیتاست‌های مورد استفاده در پیاده‌سازی ۴۱
۶.۹.۳ ساختار کد و اسکریپت‌های اصلی ۴۲
۷.۹.۳ ملاحظات عملی و محدودیت‌ها ۴۳
۸.۹.۳ جمع‌بندی ۴۴
۱۰.۳ جمع‌بندی فصل ۴۴

فصل ۴: نتایج ۴۷
۱.۴ مقدمه ۴۷
۲.۴ سوالات تحقیق و فرضیات تجربی ۴۸
۱.۲.۴ بازگویی سوالات تحقیق ۴۸
۲.۲.۴ فرضیات تجربی ۴۹
۳.۲.۴ ارتباط فرضیات با بخش‌های فصل ۴۹
۳.۴ تنظیمات آزمایش و پروتکل ارزیابی ۵۰
۱.۳.۴ دیتاست‌های مورد استفاده ۵۰

۲.۳.۴ نوع محیط‌ها و سناریوهای ارزیابی ۵۱
۳.۳.۴ معیارهای ارزیابی ۵۱
۱.۳.۳.۴ معیارهای غیرمتریک ۵۱
۲.۳.۳.۴ معیارهای متریک ۵۱
۴.۳.۴ آزمون‌های آماری ۵۲
۴.۴ نتایج کمی: مقایسهٔ عددی مدل‌ها ۵۲
۱.۴.۴ عملکرد مدل پایه در دیتاست‌های مختلف ۵۲
۲.۴.۴ مقایسهٔ مدل پایه و مدل پیشنهادی ۵۳
۳.۴.۴ تحلیل پایداری بین‌دیتاستی مدل پیشنهادی ۵۳
۵.۴ تحلیل حساسیت نسبت به پارامترهای دوربین ۵۴
۱.۵.۴ تأثیر تغییر ماتریس K بر عملکرد مدل‌ها ۵۵
۲.۵.۴ رفتار متفاوت معیارهای AbsRel و RMSE ۵۵
۳.۵.۴ ارتباط با تحلیل هندسی فصل سوم ۵۶
۶.۴ نتایج کیفی: تحلیل بصری ۵۶
۱.۶.۴ مقایسهٔ بصری نقشه‌های عمق ۵۶
۲.۶.۴ تحلیل مقیاس اجسام ۵۷
۳.۶.۴ یکنواختی سطوح و ساختار صحنه ۵۸
۴.۶.۴ بازنمایی نواحی دوردست ۵۸
۵.۶.۴ مقایسه در محیط‌های داخلی و خارجی ۵۸
۷.۴ تحلیل موردی دیتاست CityScapes ۵۹
۱.۷.۴ ماهیت و ویژگی‌های خاص دیتاست CityScapes ۵۹
۲.۷.۴ تحلیل کیفیت عمق مرجع ۵۹
۳.۷.۴ دلیل رفتار متفاوت خطاها ۵۹
۴.۷.۴ محدودیت همسان‌سازی نتایج با سایر دیتاست‌ها ۶۰
۸.۴ بحث و تفسیر نتایج ۶۰
۱.۸.۴ جمع‌بندی تحلیلی نتایج کمی و کیفی ۶۰

۲.۸.۴ پاسخ به سوالات تحقیق ۶۱
۳.۸.۴ مقایسهٔ غیرمستقیم با ادبیات موضوع ۶۱
۴.۸.۴ تفسیر نتایج در چارچوب هندسهٔ تصویربرداری ۶۲
۹.۴ جمع‌بندی فصل نتایج ۶۲

فصل ۵: بحث و نتیجه‌گیری ۶۵
۱.۵ مقدمه ۶۵
۲.۵ جمع‌بندی یافته‌های اصلی پژوهش ۶۶
۳.۵ تفسیر نهایی نتایج در چارچوب هندسهٔ تصویربرداری ۶۷
۴.۵ نوآوری‌های تحقیق ۶۸
۱.۴.۵ نوآوری تئوریک ۶۸
۲.۴.۵ نوآوری عملی ۶۸
۵.۵ محدودیت‌ها و چالش‌های پژوهش ۶۹
۶.۵ پیشنهادهایی برای تحقیقات آینده ۷۰
۷.۵ جمع‌بندی نهایی فصل ۷۱

کتاب‌نامه ۷۳

واژه‌نامهٔ فارسی به انگلیسی **اول**

واژه‌نامهٔ انگلیسی به فارسی **سوم**

فهرست تصاویر

۱.۱ مفهوم ابهام مقیاس (Scale Ambiguity): یک جسم کوچک در فاصلهٔ نزدیک می‌تواند دقیقاً همان تصویری را روی سنسور ایجاد کند که یک جسم بزرگ در فاصلهٔ دور ایجاد می‌کند. بدون اطلاعات اضافی از پارامترهای دوربین، سیستم قادر به تشخیص مقیاس واقعی فیزیکی نیست. ۲

۲.۱ تفاوت تخمین عمق متریک در مقابل غیرمتریک. خروجی مدل‌های متداول (غیرمتریک) تنها ساختار نسبی صحنه را نشان می‌دهد در حالی که یک مدل متریک دارای مقیاس فیزیکی معنادار (بر حسب متر) است. برای تولید این تصویر لطفاً خروجی مدل پایه و مدل پیشنهادی را برای یک تصویر نمونه جایگزین کنید. ۳

۱.۲ مفهوم کلی تقطیر دانش در تخمین عمق: مدل معلم دانش ساختاری و ویژگی‌های بصری خود را به مدل شاگرد منتقل می‌کند تا مدل شاگرد ضمن یادگیری مقیاس فیزیکی، دقت ساختاری خود را حفظ کند. ۹

۲.۲ دسته‌بندی کلی روش‌های تخمین عمق تک‌دوربینه در ادبیات موضوع. ۱۰

۱.۳ مدل هندسی دوربین سوراخ‌سوزنی: نگاشت یک نقطهٔ سه‌بعدی $\mathbf{X}$ به نقطهٔ دوبعدی $\mathbf{x}$ روی صفحه تصویر با توجه به فاصلهٔ کانونی f. ۲۱

۲.۳ معماری مدل پیشنهادی: اطلاعات تصویر توسط انکودر از پیش‌آموزش‌دیده استخراج می‌شود. ماتریس مشخصهٔ دوربین K پس از نرمال‌سازی توسط یک شبکهٔ تمام‌متصل به فضای نهان برده شده و به‌عنوان یک شرط هندسی به دیکودر تزریق می‌گردد. ۳۰

۱.۴ مقایسه عملکرد متریک (معیار SiLog) بین مدل پایه و پیشنهادی در دیتاست‌های مختلف. همان‌طور که مشاهده می‌شود مدل پیشنهادی در تمامی موارد خطای بسیار کمتری دارد. . . ۵۴

۲.۴ نمودار نشان‌دهنده میزان حساسیت مدل‌ها نسبت به تغییر پارامترهای کانونی دوربین. مدل پایه در مواجهه با تغییرات دارای نوسانات شدید است در حالی که مدل پیشنهادی ثبات بسیار بالاتری را نشان می‌دهد. (توجه: برای واقعی‌سازی نمودار، لطفا داده‌های حاصل از آزمون‌های خود را جایگزین مقادیر پیش‌فرض نمایید.) ۵۵

۳.۴ مقایسه کیفی نقشه‌های عمق. ستون اول: تصویر ورودی، ستون دوم: عمق مرجع، ستون سوم: خروجی مدل پایه، ستون چهارم: خروجی مدل پیشنهادی. (لطفاً شبکه‌ای از تصاویر نمونه را تولید کرده و جایگزین این تصویر نمایید.) ۵۷

۱.۵ نمای کلی از فرآیند پژوهش: حرکت از شناسایی یک محدودیت ذاتی در مدل‌های موجود تا ارائه راهکار هندسه‌آگاه و در نهایت دستیابی به پایداری متریک در تخمین عمق. ۶۷

فهرست جداول

۱.۴ مقایسهٔ معیار SiLog مدل پایه Depth Anything V2 در دیتاست‌های مختلف ۵۲
۲.۴ مقایسهٔ عملکرد مدل پایه و مدل پیشنهادی بر اساس معیار SiLog ۵۳

فهرست الگوریتم‌ها

۱.۳ روند آموزش مدل با پارامترهای هندسی ۳۴

فهرست برنامه‌ها

فهرست اختصارات

فصل ۱

مقدمه

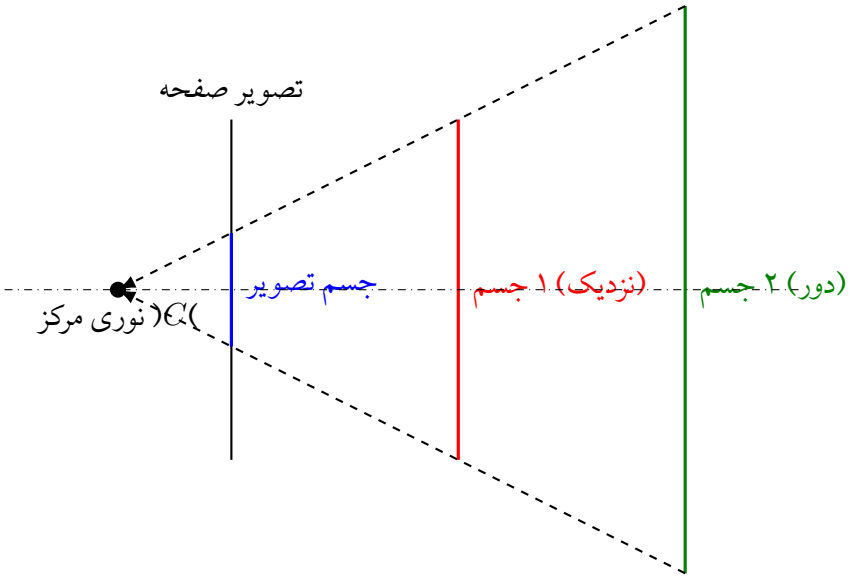
۱.۱ بیان مسئله و اهمیت آن

تخمین عمق صحنه از روی یک تصویر تک‌دوربینه یکی از مسائل بنیادی و در عین حال چالش‌برانگیز در بینایی ماشین به شمار می‌رود [۱، ۲]. این مسئله در کاربردهای متنوعی از جمله رباتیک متحرک، ناوبری خودکار، بازسازی سه‌بعدی، واقعیت افزوده و سیستم‌های کمک‌راننده نقش کلیدی ایفا می‌کند [۳]. برخلاف روش‌های مبتنی بر چند نما یا حسگرهای فعال، تخمین عمق تک‌دوربینه 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 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.

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



شکل ۱.۱: مفهوم ابهام مقیاس (Scale Ambiguity): یک جسم کوچک در فاصلهٔ نزدیک می‌تواند دقیقاً همان تصویری را روی سنسور ایجاد کند که یک جسم بزرگ در فاصلهٔ دور ایجاد می‌کند. بدون اطلاعات اضافی از پارامترهای دوربین، سیستم قادر به تشخیص مقیاس واقعی فیزیکی نیست.

منجر به تغییرات معنادار در کیفیت تخمین عمق شود.

از این منظر، دستیابی به تخمین عمق متریک 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.

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.

شکل ۲.۱: تفاوت تخمین عمق متریک در مقابل غیرمتریک. خروجی مدل‌های متداول (غیرمتریک) تنها ساختار نسبی صحنه را نشان می‌دهد در حالی که یک مدل متریک دارای مقیاس فیزیکی معنادار (بر حسب متر) است. برای تولید این تصویر لطفاً خروجی مدل پایه و مدل پیشنهادی را برای یک تصویر نمونه جایگزین کنید.

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، در عمل، تغییر نوع دوربین، فاصلهٔ کانونی یا میدان دید 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 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 مطالعات، مسئلهٔ مقیاس فیزیکی عمق یا به‌صورت ضمنی نادیده گرفته شده و یا به‌عنوان یک چالش ثانویه تلقی شده است.

از سوی دیگر، رروش‌هایی 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 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 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؟

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.

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.

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.

فصل ۲

مرور ادبیات

۱.۲ مفاهیم و تعاریف

در این بخش، به تبیین مفاهیم پایه، تعاریف فنی و مبانی نظری پرداخته می‌شود که برای درک دقیق مسئله تخمین عمق و نوآوری‌های مطرح‌شده در این پژوهش ضروری هستند.

۱.۱.۲ تخمین عمق تک‌دوربینه

تخمین عمق تک‌دوربینه[1] به فرآیند استخراج اطلاعات سه‌بعدی و فاصله نقاط صحنه از مرکز نوری دوربین، تنها با استفاده از یک تصویر دوبعدی منفرد اطلاق می‌شود [۲]. برخلاف سیستم‌های استریو که از اختلاف منظر میان دو تصویر برای محاسبه عمق استفاده می‌کنند، این رویکرد به دلیل اتکا به یک نمای واحد، ذاتاً یک مسئله «بدحال»[2] محسوب می‌شود؛ زیرا بی‌نهایت صحنه سه‌بعدی متفاوت می‌توانند منجر به ایجاد یک تصویر دوبعدی یکسان شوند [۴]. مدل‌های مدرن در این حوزه، از ویژگی‌های سطح بالا نظیر پرسپکتیو، بافت و اندازه نسبی اشیا برای غلبه بر این ابهام هندسی بهره می‌برند.

---

[1] Monocular Depth Estimation
[2] Ill-posed Problem

۲.۱.۲ عمق متریک در مقابل عمق غیرمتریک

یکی از چالش‌های بنیادین در بینایی ماشین، تمایز میان عمق نسبی (غیرمتریک[1]) و عمق مطلق (متریک[2]) است. در تخمین عمق غیرمتریک، هدف بازیابی ترتیب قرارگیری اشیا و ساختار کلی صحنه است، بدون آنکه فواصل با واحدهای فیزیکی (مانند متر) مشخص باشند. در مقابل، عمق متریک به معنای بازیابی فواصل واقعی فیزیکی است که برای کاربردهای عملیاتی نظیر ناوبری رباتیک و سیستم‌های کمک‌راننده حیاتی است [۶]. بسیاری از مدل‌های بزرگ مقیاس فعلی به دلیل آموزش بر روی داده‌های متنوع و ناهمگن، خروجی‌های غیرمتریک تولید می‌کنند که فاقد مقیاس فیزیکی پایدار هستند.

۳.۱.۲ ابهام مقیاس

ابهام مقیاس[3] به ناتوانی مدل در تشخیص اندازه واقعی اشیا و فاصله آن‌ها از دوربین بدون داشتن اطلاعات اضافی اطلاق می‌شود. در تصویربرداری تک‌دوربینه، یک جسم کوچک در فاصله نزدیک و یک جسم بزرگ در فاصله دور می‌توانند تصویری مشابه روی حسگر ایجاد کنند. این پدیده باعث می‌شود که مدل‌های یادگیری عمیق در بازیابی مقیاس فیزیکی دچار خطا شوند، مگر آنکه اطلاعاتی درباره هندسه دوربین یا ابعاد شناخته‌شده در صحنه در اختیار داشته باشند [۷].

۴.۱.۲ پارامترهای درونی دوربین

پارامترهای درونی دوربین[4] مجموعه‌ای از ویژگی‌های هندسی و اپتیکی هستند که نحوه نگاشت نقاط از فضای سه‌بعدی به صفحه تصویر را تعیین می‌کنند. این پارامترها معمولاً در قالب یک ماتریس مشخصه[5] (ماتریس K) تعریف می‌شوند که شامل فاصله کانونی[6]، مرکز تصویر[7] و ضریب پیچش است [۳]. ادغام صریح این پارامترها در مدل‌های یادگیری عمیق، امکان تبدیل اطلاعات تصویری به مختصات هندسی واقعی را فراهم کرده و نقشی کلیدی در گذار از تخمین عمق نسبی به متریک ایفا می‌کند.

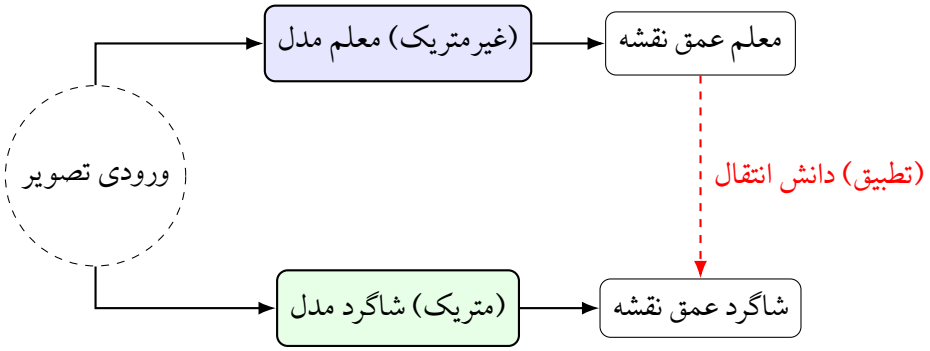
[1] Non-metric Depth
[2] Metric Depth
[3] Scale Ambiguity
[4] Camera Intrinsics
[5] Intrinsic Matrix
[6] Focal Length
[7] Principal Point

۵.۱.۲ ترنسفورمر بینایی

ترنسفورمر بینایی[1] معماری پیشرفته‌ای در یادگیری عمیق است که از مکانیسم‌های خودتوجهی[2] برای پردازش تصاویر استفاده می‌کند [۱۰]. برخلاف شبکه‌های عصبی کانولوشنی که بر ویژگی‌های محلی تمرکز دارند، ترنسفورمرها قادر به مدل‌سازی وابستگی‌های طولانی‌دامنه در کل تصویر هستند [۶]. در مدل پایه این پژوهش، یعنی Depth Anything V2، از این معماری برای استخراج ویژگی‌های غنی و مقاوم در برابر تغییرات محیطی استفاده شده است که منجر به تعمیم‌پذیری بالای مدل در سناریوهای صفرنمونه[3] می‌شود [۱۱].

۶.۱.۲ تقطیر دانش

تقطیر دانش[4] یک راهبرد آموزشی است که در آن یک مدل بزرگ و پیچیده (مدل معلم[5])، دانش خود را به یک مدل دیگر (مدل شاگرد[6]) منتقل می‌کند [۱۲]. در این پژوهش، از این رویکرد برای انتقال دانش ساختاری و ادراکی از یک مدل تخمین عمق غیرمتریک قدرتمند به مدلی استفاده شده است که با هدف یادگیری مقیاس فیزیکی و ادغام اطلاعات هندسی طراحی شده است. این فرآیند باعث می‌شود مدل شاگرد علاوه بر حفظ دقت در جزئیات ساختاری، پایداری بیشتری در تخمین مقیاس متریک داشته باشد.



شکل ۱.۲: مفهوم کلی تقطیر دانش در تخمین عمق: مدل معلم دانش ساختاری و ویژگی‌های بصری خود را به مدل شاگرد منتقل می‌کند تا مدل شاگرد ضمن یادگیری مقیاس فیزیکی، دقت ساختاری خود را حفظ کند.

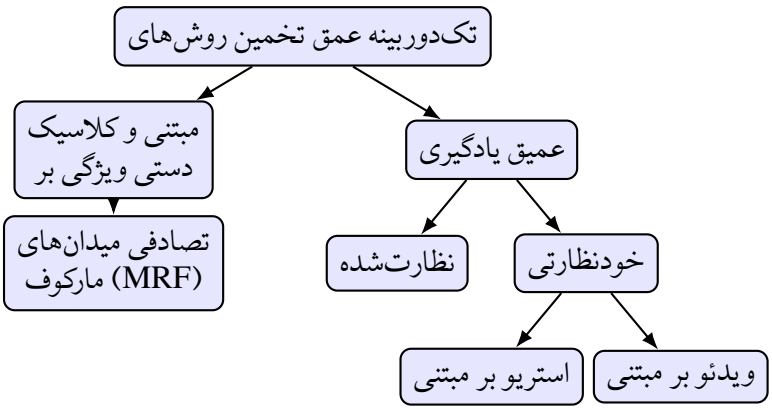
[1] Vision Transformer (ViT)
[2] Self-attention
[3] Zero-shot Generalization
[4] Knowledge Distillation
[5] Teacher Model
[6] Student Model

۷.۱.۲ یادگیری خودنظارتی

یادگیری خودنظارتی[1] روشی از یادگیری است که در آن مدل بدون نیاز به برچسب‌های دستی (نقشه‌های عمق مرجع)، و تنها با استفاده از ساختار درونی داده‌ها آموزش می‌بیند. در تخمین عمق، این امر معمولاً از طریق بازسازی دیدگاه[2] و استفاده از هندسه اپی‌پولار در ویدئوها یا تصاویر استریو محقق می‌شود [۹]. این رویکرد امکان استفاده از حجم عظیمی از داده‌های بدون برچسب را فراهم کرده و وابستگی به سنسورهای گران‌قیمت نظیر لیدار را کاهش می‌دهد.

۲.۲ مرور ادبیات و پیشینه پژوهش

پژوهش‌های انجام شده در حوزه تخمین عمق تک‌دوربینه در طول دو دهه اخیر، گذار از مدل‌های آماری ساده به شبکه‌های عصبی بسیار پیچیده را تجربه کرده‌اند. در این بخش، این مطالعات بر اساس ماهیت روش‌شناختی و نحوه مواجهه با چالش‌های هندسی دسته‌بندی و تحلیل می‌شوند.



شکل ۲.۲: دسته‌بندی کلی روش‌های تخمین عمق تک‌دوربینه در ادبیات موضوع.

[1] Self-supervised Learning
[2] View Synthesis

۱.۲.۲ رویکردهای کلاسیک و یادگیری مبتنی بر ویژگی‌های دستی

تلاش‌های اولیه برای استخراج عمق از تصاویر منفرد، عمدتاً بر نشانه‌های پرسپکتیو[1] و ویژگی‌های هندسی صریح متکی بودند. یکی از کارهای پیشگام در این حوزه، سیستم Make3D توسط [۱] ارائه شد که از یک مدل میدان تصادفی مارکوف[2] برای مدل‌سازی روابط مکانی میان قطعات تصویر استفاده می‌کرد [۱]. اگرچه این روش‌ها در بازسازی ساختارهای ساده موفق بودند، اما به دلیل وابستگی شدید به ویژگی‌های دستی[3]، در محیط‌های پیچیده و متنوع با شکست مواجه می‌شدند. در واقع، ناتوانی این مدل‌ها در استخراج مفاهیم معنایی سطح بالا از تصویر، منجر به محدودیت شدید در تعمیم‌پذیری آن‌ها می‌گردید.

۲.۲.۲ تحول در تخمین عمق با یادگیری عمیق نظارت‌شده

با ظهور شبکه‌های عصبی عمیق، رویکرد یادگیری نظارت‌شده به استاندارد طلایی این حوزه تبدیل شد. [۲] برای نخستین بار نشان دادند که استفاده از یک معماری دو مقیاسه شامل شبکه‌های کانولوشنی می‌تواند ساختار کلی و جزئیات محلی عمق را به‌طور هم‌زمان بیاموزد [۲]. در ادامه، این تیم تحقیقاتی با ترکیب ویژگی‌های معنایی و عمق، دقت مدل را بهبود بخشیدند [۵]. با این حال، همان‌طور که در پژوهش‌های بعدی اشاره شد، این مدل‌ها به دلیل آموزش بر روی دیتاست‌های محدود نظیر NYU Depth V2 [۱۳]، در مواجهه با داده‌های خارج از توزیع[4] دچار افت عملکرد جدی می‌شدند و خروجی آن‌ها به‌شدت به بافت تصویر وابسته بود تا هندسه واقعی.

۳.۲.۲ یادگیری خودنظارتی و بازسازی هندسی دیدگاه

به منظور کاهش وابستگی به داده‌های برچسب‌دار لیدار، محققان به سمت یادگیری خودنظارتی متمایل شدند. [۹] با استفاده از انسجام زمانی در ویدئوها و روابط استریو، مسئله تخمین عمق را به یک مسئله بازسازی تصویر[5] تبدیل کردند [۹]. کار آن‌ها در مدل Monodepth2 با معرفی توابع هزینه مقاوم در برابر انسداد و استفاده از شبکه‌های تخمین ژست دوربین[6] به تکامل رسید [۱۴]. با وجود این پیشرفت‌ها، یک چالش اساسی همچنان باقی ماند: ابهام مقیاس. در این روش‌ها، مدل تنها قادر به بازیابی عمق تا یک ضریب مقیاس مجهول بود که مانع

[1] Perspective Cues
[2] Markov Random Field (MRF)
[3] Hand-crafted Features
[4] Out-of-distribution Data
[5] Image Reconstruction
[6] Pose Estimation Network

از کاربرد مستقیم آن‌ها در سیستم‌های ناوبری متریک می‌شد.

۴.۲.۲ مدل‌های مبتنی بر ترنسفورمر و یادگیری در مقیاس بزرگ

در سال‌های اخیر، استفاده از معماری ترنسفورمر بینایی فصلی نو در تخمین عمق گشوده است. مدل‌هایی مانند DPT نشان دادند که جایگزینی لایه‌های کانولوشنی با مکانیسم‌های توجه جهانی می‌تواند به بازسازی مرزهای دقیق‌تر و درک بهتر از ساختار جهانی صحنه منجر شود [۶]. اخیراً مدل Depth Anything V2 با بهره‌گیری از داده‌های عظیم و تکنیک‌های پیشرفته تقطیر دانش، به پایداری بی‌سابقه‌ای در تخمین عمق غیرمتریک دست یافته است [۱۱]. با این حال، این مدل‌های قدرتمند نیز همچنان اطلاعات درونی دوربین را نادیده می‌گیرند. در حالی که [۶] بر تعمیم‌پذیری در محیط‌های مختلف تأید داشتند، نادیده گرفتن پارامترهای هندسی باعث می‌شود که خروجی این مدل‌ها با تغییر فاصله کانونی دوربین دچار نوسانات مقیاسی شود.

۵.۲.۲ ادغام دانش هندسی و چالش پایداری متریک

تلاش‌های محدودی برای پیوند میان هندسه کلاسیک و یادگیری عمیق صورت گرفته است. [۷] نشان دادند که بدون داشتن مقیاس پایه، تخمین عمق مطلق غیرممکن است [۷]. برخی پژوهش‌ها نظیر کار [۸] تلاش کردند با استفاده از ویژگی‌های معنایی هدایت‌شده توسط هندسه، مقیاس را بازیابی کنند [۸]. با این حال، خلأ موجود در ادبیات فعلی، نبود یک راهکار نظام‌مند برای تزریق مستقیم ماتریس مشخصه دوربین به دیکودرهای مدل‌های بزرگ‌مقیاس است. اکثر روش‌های فعلی یا بر کالیبراسیون پس‌پردازشی تکیه دارند و یا فرض می‌کنند شبکه به‌صورت ضمنی هندسه را می‌آموزد؛ فرضی که در شرایط تغییر دوربین به چالش کشیده می‌شود. پژوهش حاضر بر ادغام صریح پارامترهای درونی در معماری Depth Anything V2 تمرکز دارد.

۳.۲ شکاف تحقیقاتی

با وجود پیشرفت‌های چشمگیر در حوزه تخمین عمق تک‌دوربینه، تحلیل انتقادی ادبیات موضوع نشان‌دهنده وجود خلأهای تحقیقاتی جدی است که مانع از کاربرد گسترده این فناوری در محیط‌های عملیاتی می‌شود.

۱.۳.۲ فقدان پایداری در تخمین عمق متریک

بخش عمده‌ای از پژوهش‌های اخیر بر بهبود دقت نسبی معطوف شده‌اند [۱۱، ۶]. با این حال، این مدل‌ها در تولید خروجی با مقیاس فیزیکی واقعی دچار ضعف هستند. شکاف موجود، ناتوانی مدل‌ها در حفظ پایداری مقیاس[1] در مواجهه با صحنه‌هایی است که از نظر بصری مشابه اما از نظر هندسی متفاوت هستند. فقدان یک مکانیسم درونی برای تبدیل ویژگی‌های بصری به مقادیر فیزیکی، باعث می‌شود که این مدل‌ها بدون کالیبراسیون خارجی در کاربردهای متریک غیرقابل استفاده باقی بمانند.

۲.۳.۲ نادیده گرفتن صریح هندسه و پارامترهای درونی دوربین

اکثر مدل‌های فعلی فرض می‌کنند که شبکه عصبی می‌تواند پارامترهای درونی دوربین را به‌صورت ضمنی بیاموزد. با این حال، تغییر در مشخصات سنسور یا لنز منجر به بروز خطا در تخمین عمق می‌شود [۷]. خلأ اصلی تحقیقاتی، نبودِ یک معماری یکپارچه است که بتواند ماتریس مشخصه دوربین را به‌عنوان یک ورودی مرتبه اول پردازش کند. اکثر روش‌ها یا هندسه را نادیده می‌گیرند و یا از آن تنها در مراحل پس‌پردازش استفاده می‌کنند که منجر به عدم انطباق ویژگی‌ها با قیود هندسی در لایه‌های میانی شبکه می‌شود.

۳.۳.۲ حساسیت بالا نسبت به تغییرات دامنه و داده‌های خارج از توزیع

مدل‌هایی که بر روی دیتاست‌های خاص آموزش دیده‌اند، در مواجهه با داده‌های خارج از توزیع دچار پدیده لغزش مقیاس[2] می‌شوند [۱۵]. اگرچه روش‌های تقطیر دانش تلاش کرده‌اند این مشکل را حل کنند [۱۱]، اما وابستگی خروجی متریک به توزیع داده‌های آموزشی همچنان یک چالش است. شکاف تحقیقاتی در این است که پایداری متریک باید بر اساس درک صریح از رابطه فیزیکی پیکسل‌ها و دوربین باشد، نه صرفاً حافظه آماری از داده‌ها.

۴.۳.۲ محدودیت معیارهای ارزیابی در تفکیک خطاهای هندسی

در بسیاری از مطالعات، از معیارهای نرمال‌شده‌ای نظیر AbsRel استفاده می‌شود که قادر به آشکارسازی خطاهای ناشی از عدم پایداری مقیاس فیزیکی نیستند [۶]. خلأ موجود، کمبود تحلیل‌های عمیق بر روی

[1] Scale Stability
[2] Scale Drift

معیارهای حساس به واحد فیزیکی (مانند RMSE) در سناریوهایی است که پارامترهای درونی دوربین تغییر می‌کنند.

۴.۲ نوآوری‌های تحقیق

پژوهش حاضر با هدف گذار به تخمین عمق هندسه‌آگاه[1]، چندین نوآوری کلیدی را ارائه می‌دهد:

۱.۴.۲ ادغام صریح پارامترهای درونی دوربین در معماری ترنسفورمر

نوآوری اصلی این پژوهش، تزریق مستقیم و صریح ماتریس مشخصه دوربین به درون ساختار دیکودر مدل Depth Anything V2 است. در این رویکرد، پارامترهای هندسی پس از نرمال‌سازی و نگاشت به یک فضای نهان[2]، به‌عنوان اطلاعات مکمل در کنار ویژگی‌های بصری قرار می‌گیرند. این نوآوری مستقیماً شکاف مربوط به ناپایداری مقیاس را هدف قرار می‌دهد.

۲.۴.۲ طراحی دیکودر چندمؤلفه‌ای با قابلیت تفکیک ویژگی‌های بصری و هندسی

در این پژوهش، معماری دیکودر به‌گونه‌ای اصلاح شده است که با استفاده از مکانیزم‌های ادغام ویژگی[3]، وزن دهی به ویژگی‌های بصری را بر اساس پارامترهای درونی دوربین تنظیم کند. این ساختار اجازه می‌دهد که مدل در سناریوهای صفرنمونه، پایداری متریک خود را حتی در صورت تغییرات شدید در فاصله کانونی حفظ نماید.

۳.۴.۲ راهبرد آموزش چندمرحله‌ای مبتنی بر تقطیر دانش متریک

این تحقیق یک چارچوب آموزشی جدید پیشنهاد می‌دهد که در آن از تقطیر دانش [۱۲] برای انتقال پایداریِ ادراکی از یک مدل معلم غیرمتریک به یک مدل شاگرد متریک استفاده می‌شود. نوآوری در طراحی تابع هزینه

[1] Geometry-aware Depth Estimation
[2] Latent Space
[3] Feature Fusion Mechanisms

ترکیبی است که هم‌زمان بر حفظ ساختار لبه‌ها و انطباق مقیاس فیزیکی با داده‌های واقعیت‌میدانی[1] تمرکز دارد.

۴.۴.۲ تحلیل حساسیت سیستماتیک نسبت به تغییرات هندسه تصویربرداری

این پژوهش با ارائه یک چارچوب ارزیابی دقیق، تأثیر تغییر پارامترهای دوربین را بر معیارهای مختلف بررسی می‌کند. این نوآوری تحلیلی نشان می‌دهد که چگونه معیارهای سنتی ممکن است در نمایش خطاهای مقیاس ناتوان باشند و در مقابل، معیارهایی نظیر RMSE و SiLog رفتار هندسی واقعی مدل را منعکس می‌کنند [۱۴].

۵.۲ جمع‌بندی فصل

در این فصل، با بررسی پیشینه پژوهش مشخص گردید که مدل‌های فعلی علی‌رغم دقت بصری بالا، در حفظ «پایداری مقیاس متریک» دچار ضعف هستند. این چالش ریشه در نادیده گرفتن صریح پارامترهای هندسی دوربین دارد که منجر به ابهام مقیاس در محیط‌های متنوع می‌شود. پژوهش حاضر با هدف رفع این شکاف، بر ادغام مستقیم ماتریس مشخصه دوربین در معماری مدل Depth Anything V2 تمرکز دارد. در فصل آینده، جزئیات متدولوژی، فرمول‌بندی ریاضی این ادغام و راهبرد آموزش مدل پیشنهادی به‌صورت تفصیلی تشریح خواهد شد.

[1] Ground Truth

فصل ۳

روش تحقیق

۱.۳ مقدمه

در این فصل، روش پیشنهادی این پژوهش برای تخمین عمق تک‌چشمی بدون نظارت تشریح می‌شود. تمرکز اصلی این فصل بر توضیح دقیق معماری مدل، نحوه ادغام اطلاعات هندسی دوربین در شبکه عصبی، و فرآیند پیاده‌سازی و آموزش سیستم است. با توجه به مباحث مطرح‌شده در فصل‌های پیشین، فرض بر این است که خواننده با مفاهیم کلی تخمین عمق، یادگیری بدون نظارت و معماری‌های مبتنی بر Transformer آشناست؛ بنابراین، از تکرار مباحث مقدماتی پرهیز شده و توضیحات مستقیماً بر جنبه‌های روش‌شناختی متمرکز هستند.

مدل پایه مورد استفاده در این پژوهش، Depth Anything V2 است که یک معماری encoder-decoder مبتنی بر Transformer Vision را برای تخمین عمق چگال از تصاویر تک‌چشمی ارائه می‌دهد [۱۱]. 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 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 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 بدون نظارت

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، برخلاف روش‌های نظارت‌شده، هیچ‌گونه برچسب عمق صریحی در فرآیند آموزش در دسترس نیست. در عوض، یادگیری مبتنی بر قیود هندسی، سازگاری تصویری، و روابط بازسازی میان نماهای مختلف صورت می‌گیرد. در رایج‌ترین تنظیمات، دادهٔ آموزشی شامل توالی‌هایی از تصاویر متوالی یک صحنه است که با یک دوربین متحرک ثبت شده‌اند [۷، ۹]. در این حالت، مدل با بهره‌گیری از اصل سازگاری فوتومتریک، می‌آموزد که نگاشت عمق پیش‌بینی‌شده باید به‌گونه‌ای باشد که امکان بازسازی یک تصویر از روی تصویر دیگر فراهم شود.

[1] Monocular Depth Estimation

به‌صورت صوری، اگر I_t تصویر ورودی در زمان t باشد، هدف یادگیری تابعی است که نگاشت عمق D_t را 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، مسئلهٔ تخمین عمق تک‌دوربینه بدون نظارت به‌صورت زیر بازتعریف می‌شود: هدف، یادگیری تابعی است که علاوه بر تصویر ورودی، اطلاعات مربوط به پارامترهای درونی دوربین را نیز دریافت کرده و نگاشت عمقی را تولید کند که از نظر هندسی با مدل تصویربرداری دوربین سازگار باشد. به‌صورت نمادین، نگاشت مورد نظر به‌شکل زیر بیان می‌شود:

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

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.

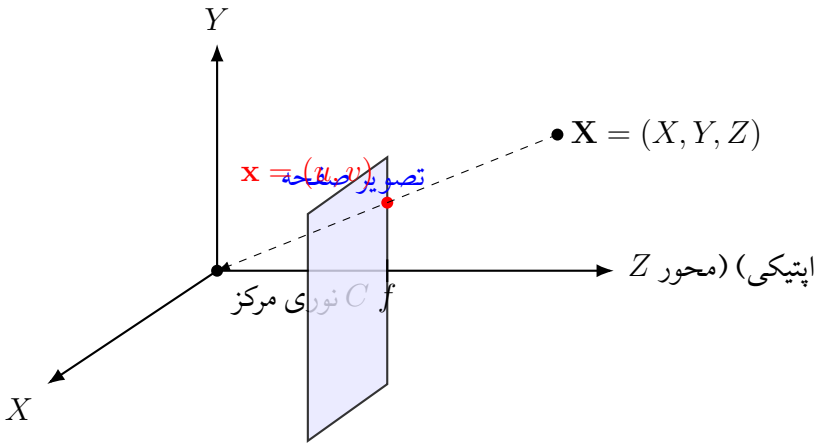
در این بخش، ابتدا مدل هندسی دوربین مورد استفاده معرفی می‌شود، سپس ماتریس مشخصهٔ دوربین به‌صورت صوری تعریف شده و در ادامه، تأثیر مستقیم پارامترهای درونی بر مقیاس و پایداری تخمین عمق مورد بررسی قرار می‌گیرد.

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



شکل ۱.۳: مدل هندسی دوربین سوراخ‌سوزنی: نگاشت یک نقطهٔ سه‌بعدی $\mathbf{X}$ به نقطهٔ دوبعدی $\mathbf{x}$ روی صفحه تصویر با توجه به فاصلهٔ کانونی f.

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 [۱۱]. 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 [۱۰]. 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[۱] شده و وابستگی‌های غیرمحلی میان

[۱] 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.

بخش‌های مختلف تصویر را مدل‌سازی می‌کنند. استفاده از سازوکار [1]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 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.

[1] Attention

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

تحقیقات اخیر، از جمله تحلیل‌های ارائه‌شده در فصل بعد این پژوهش ، به‌روشنی نشان داده‌اند که تفاوت در پارامترهای دوربین، علی‌رغم شباهت ظاهری تصاویر، منجر به بروز اختلاف معنادار در دقت تخمین عمق بر اساس معیارهای متریک[3] می‌شود. برای نمونه، با محاسبه و مقایسهٔ شاخص‌های خطای مطلق نسبی[4] و ریشهٔ میانگین مربعات[5]، مشخص شد که گرچه معیار AbsRel نسبت به تغییر مشخصه‌های دوربین حساسیت چشمگیری نشان نمی‌دهد (اختلاف آماری غیرمعنادار)، اما که—RMSE به مقیاس فیزیکی (متریک) عمق وابسته است—تحت تأثیر اختلافات ظریف میان دوربین‌ها به‌صورت معناداری افزایش می‌یابد.

برای مثال، در آزمایش حاضر، داده‌های مجموعهٔ ،Middlebury با وجود استفاده از دو دوربین با تفاوت‌های ظریف خصوصاً در مؤلفهٔ f_y، به‌طور مستقیم منجر به اختلاف معنادار در میزان RMSE تخمین عمق شدند. تحلیل آزمون t-test نشان داد که مقدار p برای اختلاف RMSE میان دوربین‌های مختلف، کمتر از بوده است، که بیانگر وجود تفاوت معنادار آماری بین نتایج مدل روی داده‌های ثبت‌شده با مشخصه‌های فنی متفاوت است. این در حالی است که همان آزمایش روی AbsRel چنین حساسیت معناداری را بروز نداده است.

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

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 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 [۱۶]، بر همین نکته تأکید دارند: لحاظ صریح مشخصات دوربین، کلید دستیابی به یک تخمین عمق متریک باکیفیت و مقاوم در کاربردهای واقعی است.

[1] Camera Intrinsic Matrix (K)

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، ۱۵].

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 پایه 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 [۱۷، ۱۸].

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 [۱۹، ۷]. به همین دلیل، در روش پیشنهادی، پارامترهای درونی دوربین پیش از ورود به شبکه نرمال‌سازی می‌شوند. به‌طور خاص، مؤلفه‌های فاصلهٔ کانونی بر حسب ابعاد تصویر نرمال شده و مختصات نقطهٔ اصلی نسبت به مرکز تصویر بازپارامتردهی می‌شوند:

$$\tilde{f}_x = \frac{f_x}{W}, \quad \tilde{f}_y = \frac{f_y}{H}, \quad \tilde{c}_x = \frac{c_x - W/۲}{W}, \quad \tilde{c}_y = \frac{c_y - H/۲}{H}, \tag{۶.۳}$$

که در آن W و H به‌ترتیب عرض و ارتفاع تصویر ورودی هستند. این بازنمایی نرمال‌شده، امکان استفادهٔ پایدار از اطلاعات دوربین در مجموعه‌داده‌هایی با رزولوشن‌های متفاوت را فراهم می‌کند [۲۰، ۲۱].

در نهایت، بردار پارامترهای درونی دوربین به‌صورت

$$\mathbf{k} = [\tilde{f}_x, \tilde{f}_y, \tilde{c}_x, \tilde{c}_y] \tag{۷.۳}$$

تعریف می‌شود و به‌عنوان ورودی هندسی مدل مورد استفاده قرار می‌گیرد.

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 ماژول نگاشت نهان استفاده می‌شود که وظیفهٔ آن، تبدیل بردار کم‌بعد $\mathbf{k}$ به یک بازنمایی نهان سازگار با فضای ویژگی‌های شبکه است. این ماژول شامل چند لایهٔ کاملاً متصل است که خروجی آن به یک بردار نهان با ابعاد d_k نگاشت می‌شود:

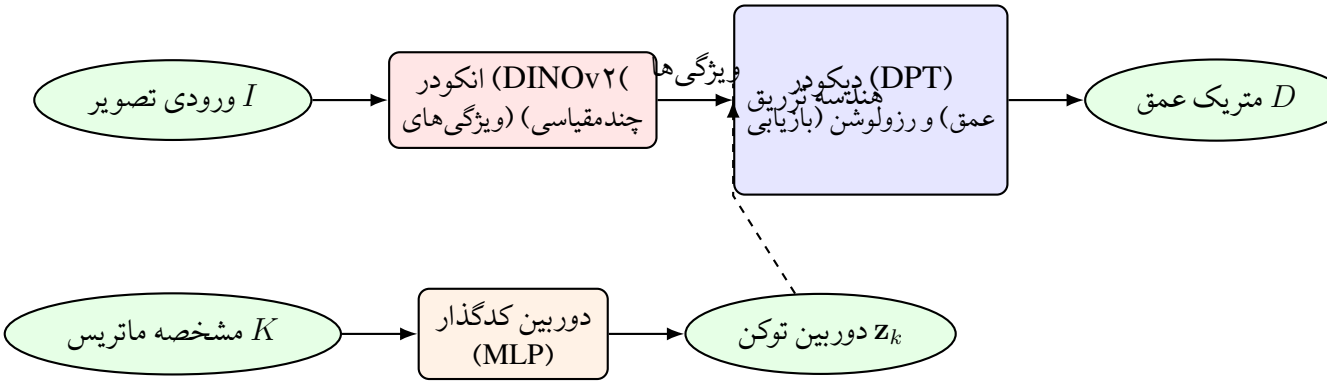
$$\mathbf{z}_k = \phi(\mathbf{k}), \tag{۸.۳}$$

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

بازمایی $\mathbf{z}_k$ سپس در نقاط مشخصی از معماری encoder–decoder به شبکه تزریق می‌شود. در این پژوهش، تمرکز اصلی بر ادغام این اطلاعات در بخش decoder است، جایی که فرآیند بازسازی نقشهٔ عمق انجام می‌شود و حساسیت مدل نسبت به مقیاس بیشترین مقدار را دارد [۶، ۲۲]. به‌طور خاص، $\mathbf{z}_k$ پس از بسط مکانی، با ویژگی‌های چندمقیاسی decoder ترکیب می‌شود تا اطلاعات هندسی دوربین مستقیماً بر فرآیند تخمین عمق اثر بگذارد. جزئیات دقیق محل و نحوهٔ این ادغام، با الهام از مطالعات پیشین در معماری‌های encoder-decoder برای پیش‌بینی چگال، به‌گونه‌ای انتخاب شده است که اطلاعات هندسی در مرحله‌ای وارد شبکه شود که حساسیت نسبت به مقیاس بیشینه است. مطابق با تحلیل‌های ارائه‌شده در معماری DPT، تزریق اطلاعات کمکی در مراحل ابتدایی decoder، پس از استخراج بازنمایی‌های سطح‌بالا و پیش از بازسازی نهایی وضوح مکانی، منجر به پایداری بیشتر در تخمین کمیت‌های متریک می‌شود [۶].

به‌طور مشخص، پژوهش‌های مبتنی بر تنظیم ویژگی‌ها به‌کمک بردارهای شرطی، نظیر روش‌های مبتنی بر Modulation Linear Feature-wise (FiLM)، نشان داده‌اند که تزریق اطلاعات هندسی یا سنسوری در قالب بردارهای نهان، در لایه‌های میانی یا ابتدایی decoder، نسبت به تزریق در encoder، اثرگذاری مستقیم‌تری بر خروجی‌های متریک دارد [۲۳]. این راهبرد، از تداخل اطلاعات هندسی با فرآیند استخراج ویژگی‌های معنایی جلوگیری کرده و امکان کنترل صریح مقیاس در مرحلهٔ بازسازی را فراهم می‌کند.

بر این اساس، در روش پیشنهادی، بازنمایی نهان پارامترهای درونی دوربین $\mathbf{z}_k$ پس از بسط مکانی، به‌صورت الحاق یا جمع شرطی، به ورودی اولین بلوک decoder تزریق می‌شود؛ انتخابی که با نتایج گزارش‌شده در روش‌های آگاه از هندسهٔ دوربین برای تخمین عمق و جریان صحنه هم‌راستا است [۱۷، ۲۲].



شکل ۲.۳: معماری مدل پیشنهادی: اطلاعات تصویر توسط انکودر ازپیش‌آموزش‌دیده استخراج می‌شود. ماتریس مشخصهٔ دوربین K پس از نرمال‌سازی توسط یک شبکهٔ تمام‌متصل به فضای نهان برده شده و به‌عنوان یک شرط هندسی به دیکودر تزریق می‌گردد.

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 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

در این پژوهش، آموزش مدل پیشنهادی به‌صورت فاین‌تیون[1] یک مدل ازپیش‌آموزش‌یافته انجام شده است. به‌طور خاص، از نسخهٔ non-metric مدل Depth Anything V2 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

[1] Fine-tuning

هندسی جدید را بدون نیاز به آموزش از ابتدا فراهم می‌کند.

۱.۷.۳ مجموعه‌داده‌های مورد استفاده

آموزش مدل بر اساس ترکیبی از مجموعه‌داده‌های Virtual KITTI[1]، NYUv2[2] و DIODE[3] انجام شده است. این مجموعه‌داده‌ها از نظر نوع صحنه (داخلی و خارجی)، شرایط تصویربرداری، و مشخصات دوربین، تنوع قابل توجهی دارند و از این منظر، بستر مناسبی برای ارزیابی پایداری مدل نسبت به تغییر پارامترهای درونی دوربین فراهم می‌کنند [۲۴، ۱۳، ۲۵].

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 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 مدل پایه و راهبرد فاین‌تیون

مدل پایهٔ مورد استفاده، نسخهٔ Depth Anything V2 non-metric است که به‌صورت teacher-student distillation[4] 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.

برای حفظ پایداری آموزش و جلوگیری از بیش‌برازش، بخش encoder مبتنی بر DINOv2[5] 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 [۲۶]. در مقابل،

[1] Virtual KITTI
[2] NYUv2
[3] DIODE
[4] Teacher–Student Distillation
[5] DINOv2

بخش decoder و همچنین ماژول‌های مرتبط با تخمین عمق و پردازش پارامترهای دوربین، به‌صورت کامل قابل یادگیری باقی مانده‌اند.

۳.۷.۳ فعال‌سازی پشتیبانی از پارامترهای دوربین

برای فعال‌سازی استفاده از ماتریس مشخصهٔ دوربین در فرآیند آموزش، از گزینهٔ `--use-camera-intrinsics` در اسکریپت آموزش استفاده شده است. 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 تعیین می‌کند که بازنمایی نهان پارامترهای دوربین در کدام لایه از شبکه وارد شود. در تنظیمات اولیه، تزریق در لایهٔ ابتدایی decoder انجام شده است، اما بررسی تأثیر این انتخاب بر عملکرد نهایی مدل در بخش ارزیابی تجربی مورد توجه قرار خواهد گرفت. [نیاز به آزمایش حساسیت نسبت به لایهٔ تزریق]

۴.۷.۳ تنظیمات آموزش و بهینه‌سازی

فرآیند آموزش با استفاده از اسکریپت `train.py` انجام شده است. در این اسکریپت، نوع مدل به‌صورت متریک تنظیم شده است:

$$\text{–model-type metric} \tag{۹.۳}$$

که به این معناست که خروجی شبکه، تخمین عمق مطلق بر حسب واحد متر خواهد 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 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.

۵.۷.۳ استفاده از تقطیر دانش

در فرآیند آموزش، از تقطیر دانش به‌عنوان یک قید نظارتی ضعیف استفاده شده است. در این چارچوب، مدل Depth Anything V2 non-metric به‌عنوان معلم عمل کرده و خروجی آن، به‌عنوان مرجع نسبی، مدل دانش‌آموز را در فرآیند یادگیری هدایت می‌کند. این راهبرد، به‌ویژه در نبود برچسب‌های دقیق عمق متریک برای تمام داده‌ها، نقش مهمی در انتقال دانش ساختاری از مدل معلم به مدل متریک پیشنهادی ایفا می‌کند.

فعال‌سازی این سازوکار از طریق پارامترهای `--use-distillation` و `--teacher-checkpoint` انجام شده است. به این ترتیب، مدل پیشنهادی نه‌تنها از داده‌های آموزشی، بلکه از پیش‌دانش مدل معلم نیز بهره‌مند می‌شود.

۶.۷.۳ جمع‌بندی فرآیند آموزش

در مجموع، راهبرد آموزش این پژوهش بر پایهٔ فاین‌تیون کنترل‌شدهٔ یک مدل قدرتمند ازپیش‌آموزش‌یافته، با فریز کردن بخش‌های پایدار، فعال‌سازی صریح پارامترهای دوربین، و استفاده از تقطیر دانش طراحی شده است. این ترکیب، امکان آموزش مدلی را فراهم می‌کند که ضمن حفظ توانایی‌های عمومی Depth Anything V2، قادر به ارائهٔ تخمین عمق متریک پایدار و سازگار با مشخصات واقعی دوربین باشد.

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

[1] Gradient Clipping

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 پایه 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 $\leftarrow D^{GT}$
    {استخراج ویژگی‌های بصری}
    $F \leftarrow Encoder(I, \theta_{enc})$
    {پیش‌پردازش و استخراج ویژگی‌های دوربین}
    $\mathbf{k} \leftarrow$ نرمال‌سازی(K, W, H)
    $\mathbf{z}_k \leftarrow MLP(\mathbf{k}, \theta_{cam})$
    {تخمین عمق با تزریق اطلاعات هندسی}
    $D^{pred} \leftarrow Decoder(F, \mathbf{z}_k, \theta_{dec})$
    {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

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.

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:

$$\mathcal{L}_{\text{SiLog}} = \sqrt{\frac{۱}{N}\sum_{i=۱}^{N} d_i^۲ - \frac{\lambda}{N^۲}\left(\sum_{i=۱}^{N} d_i\right)^۲}, \tag{۱۱.۳}$$

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 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[2] استفاده می‌شود. این تابع هزینه ابتدا پیش‌بینی مدل را با استفاده از روش کمترین مربعات بر مقدار واقعی منطبق می‌کند:

[1] Scale-Invariant Logarithmic Loss
[2] Scale-Shift Invariant Loss

$$D^*_{pred} = s \cdot D_{pred} + t, \tag{۱۲.۳}$$

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

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

$$\mathcal{L}_{\text{SSI}} = (۱ - \alpha)\mathcal{L}_{\text{pixel}} + \alpha\mathcal{L}_{\text{grad}}, \tag{۱۳.۳}$$

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.

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)**

$$\text{AbsRel} = \frac{۱}{N}\sum_{i=۱}^{N}\frac{|D_i^{pred} - D_i^{GT}|}{D_i^{GT}} \tag{۱۴.۳}$$

- **RMSE:**

$$\text{RMSE} = \sqrt{\frac{۱}{N}\sum_{i=۱}^{N}(D_i^{pred} - D_i^{GT})^۲} \tag{۱۵.۳}$$

- **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.

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، 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:

$$\mathrm{RMSE} = \sqrt{\frac{۱}{N}\sum_{i=۱}^{N}(D_i^{pred} - D_i^{GT})^۲}. \tag{۱۶.۳}$$

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، 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. تصاویر ابتدا به اندازهٔ ثابت ۵۱۸ × ۵۱۸ پیکسل تغییر اندازه داده شده و سپس در صورت نیاز، برش‌های تصادفی با اندازهٔ کوچک‌تر برای افزایش تنوع داده‌ای در مرحلهٔ آموزش اعمال می‌گردد. این راهبرد، مطابق با تنظیمات مدل پایه Depth Anything V2، به بهبود تعمیم‌پذیری مدل کمک می‌کند [۱۱].

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

$$I_{norm} = \frac{I - \mu_{ImageNet}}{\sigma_{ImageNet}}, \tag{۱۷.۳}$$

[1] Random Seed

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:

$$\tilde{f}_x = \frac{f_x}{W}, \quad \tilde{f}_y = \frac{f_y}{H}, \quad \tilde{c}_x = \frac{c_x}{W}, \quad \tilde{c}_y = \frac{c_y}{H}, \tag{۱۸.۳}$$

که در آن W و H به‌ترتیب عرض و ارتفاع تصویر ورودی هستند.
بردار نهایی پارامترهای دوربین به‌صورت

$$\mathbf{k} = [\tilde{f}_x, \tilde{f}_y, \tilde{c}_x, \tilde{c}_y] \tag{۱۹.۳}$$

تعریف شده و سپس از طریق یک شبکهٔ تمام‌متصل کوچک (شامل دو لایهٔ خطی با تابع فعال‌سازی ReLU) به یک نمایش نهان $\mathbf{z}_k$ با بُعد ۲۵۶ نگاشت می‌شود. این نمایش نهان در مرحلهٔ دیکودینگ به ویژگی‌های تصویری تزریق می‌گردد.

۴.۹.۳ تنظیمات آموزش و راهبرد بهینه‌سازی

مدل پیشنهادی با استفاده از بهینه‌ساز AdamW [۲۷] 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: $\beta_1 = ۰/۹$، $\beta_2 = ۰/۹۹۹$
- 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.

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 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

[1] Head Layers
[2] Gradient Accumulation

این تنوع دیتاست‌ها، به‌ویژه از منظر تفاوت در پارامترهای درونی دوربین و میدان دید، نقش مهمی در ارزیابی پایداری متریک مدل پیشنهادی ایفا می‌کند. برای هر دیتاست، فایل‌های لیست تقسیم‌بندی داده (train/validation/test splits) تهیه شده که شامل مسیر تصاویر، نقشه‌های عمق و پارامترهای درونی دوربین در قالب فایل‌های npy. هستند.

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/          # base model
          DepthAnythingV2-revised/ # model with intrinsics
          Depth-Anything/3-
      base.py                       # base classes
  datasets/
      base.py                       # base dataset class
      training_datasets.py          # training datasets
      (middlebury, cityscapes, ...)
  src/
      pipeline.py                   # main pipeline
      metrics.py                    # evaluation metrics
      visualization.py              # visualization tools
  train.py                           # main training script
  evaluate.py                        # evaluation script
  compare_models.py                  # statistical comparison
```

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 (basic metric)
- `--encoder`: انتخاب اندازهٔ انکودر (vits, vitb, vitl, vitg)
- `--use-distillation`: فعال‌سازی تقطیر دانش
- `--freeze-dinov2`: فریز کردن وزن‌های انکودر
- `--lambda-distill` و `--lambda-metric`: تنظیم ضرایب تابع هزینه

علاوه بر اسکریپت آموزش، اسکریپت‌های مستقلی برای مقایسهٔ خروجی مدل‌ها، تحلیل آماری نتایج (شامل آزمون t زوجی و محاسبهٔ فواصل اطمینان (bootstrap و تجسم نتایج پیاده‌سازی شده‌اند.

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 تکنیک‌هایی نظیر accumulation gradient و training precision mixed استفاده شده است.
- **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 نمونه (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.
- **تکرارپذیری:** علی‌رغم تلاش برای ثابت نگه‌داشتن تمامی بذرهای تصادفی، برخی عملیات GPU ممکن است غیرقطعی باشند. برای کاهش این اثر، هر آزمایش حداقل سه بار تکرار شده و میانگین نتایج گزارش شده است.

۸.۹.۳ جمع‌بندی

در این بخش، جزئیات پیاده‌سازی مدل پیشنهادی 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 تخمین عمق تک‌دوربینه بدون نظارت 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 یک مدل بنیادین یادگیری عمیق 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 و ادغام پارامترهای درونی دوربین در ساختار دیکودر متمرکز شد، به‌گونه‌ای که اطلاعات هندسی از اطلاعات صرفاً تصویری تفکیک شده و به‌صورت کنترل‌شده در فرآیند بازسازی عمق وارد شوند.

راهبرد پیشنهادی شامل نرمال‌سازی پارامترهای ماتریس مشخصهٔ دوربین، نگاشت آن‌ها به یک فضای نهان کم‌بعد، و تزریق این نمایش هندسی به دیکودر مدل بود. این طراحی، بدون ایجاد اختلال در انکودر از پیش‌آموزش‌دیده، امکان یادگیری وابستگی‌های متریک را فراهم می‌کند. علاوه بر این، راهبرد آموزش چندمرحله‌ای مبتنی بر تقطیر

دانش از یک مدل غیرمتریک، بستری مناسب برای انتقال دانش ساختاری عمق و سپس انطباق آن با مقیاس فیزیکی فراهم ساخت.

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 بدون نظارت ارائه داد. در فصل بعد، عملکرد مدل پیشنهادی به‌صورت تجربی بر روی دیتاست‌های متنوع indoor و outdoor ارزیابی شده و تأثیر افزودن ماتریس مشخصهٔ دوربین بر پایداری متریک، دقت عددی و حساسیت مدل نسبت به تغییرات پارامترهای تصویربرداری به‌طور جامع تحلیل خواهد شد.

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، مجموعه‌ای از آزمایش‌های کمی بر روی دیتاست‌های متنوع شامل محیط‌های داخلی، خارجی، واقعی و مصنوعی انجام شده است. در این آزمایش‌ها، مدل پیشنهادی با

---

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

مدل پایه 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 Depth Anything V2 در حالت پایه، 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 پایه 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، منجر به کاهش معنادار خطا در معیارهای متریک و بهبود پایداری مقیاس فیزیکی خروجی عمق می‌شود.
- **فرضیهٔ سوم:** بهبود عملکرد متریک مدل پیشنهادی، 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 پایه در دیتاست‌های مختلف بررسی می‌شود.
- فرضیهٔ دوم در بخش‌های مقایسهٔ کمی بین مدل پایه و مدل پیشنهادی و همچنین در تحلیل پایداری بین‌دیتاستی مدل پیشنهادی 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

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 (CityScapes DrivingStereo 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.

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

۴.۳.۴ آزمون‌های آماری

به‌منظور بررسی معناداری تفاوت‌های مشاهده‌شده بین مدل‌ها و دیتاست‌ها، از آزمون‌های آماری استفاده شده است. در این پژوهش، آزمون t دو نمونه‌ای[1] برای مقایسهٔ میانگین خطاها به‌کار رفته و سطح معناداری برابر با ۰٫۰۵ در نظر گرفته شده است.

علاوه بر این، به‌منظور برآورد پایداری نتایج و کاهش وابستگی به فرضیات توزیعی، از بازنمونه‌گیری بوت‌استرپ[2] برای محاسبهٔ بازه‌های اطمینان ۹۵٪ استفاده شده است. ترکیب این دو رویکرد 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 پایه Depth Anything V2 در دیتاست‌های مختلف و مقایسهٔ آن با مدل پیشنهادی 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.

۱.۴.۴ عملکرد مدل پایه در دیتاست‌های مختلف

در گام نخست، عملکرد مدل پایه 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 پایه در جدول ۱.۴ ارائه شده است.

جدول ۱.۴: مقایسهٔ معیار SiLog مدل پایه Depth Anything V2 در دیتاست‌های مختلف

مقایسهٔ دیتاست‌ها	اختلاف میانگین	p-value	نتیجه
VKITTI vs DrivingStereo	۰۳۹.۰	$<$ ۰٫۰۵	معنادار
NYUv2 vs DrivingStereo	۱۰۱.۰	$<$ ۰٫۰۵	معنادار
NYUv2 vs DIODE	۴۵۰.۰	$<$ ۰٫۰۵	معنادار

[1] Two-sample t-test
[2] Bootstrap Resampling

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

در ادامه، عملکرد مدل پایه با مدل پیشنهادی 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 پایه و مدل پیشنهادی بر اساس معیار SiLog

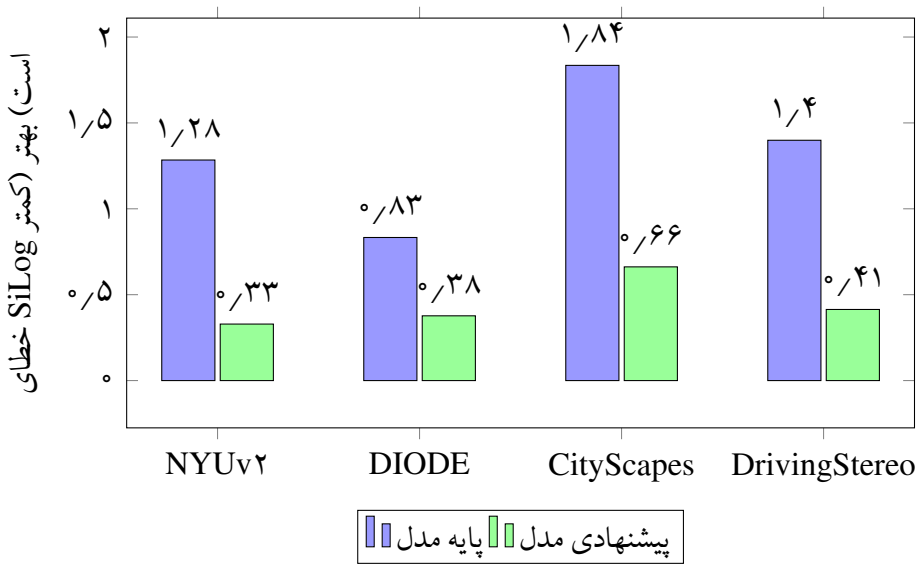
دیتاست	مدل پایه	مدل پیشنهادی	بهبود (%)	p-value
NYUv2	۲۸۴.۱	۳۲۹.۰	۳۸.۷۴	$\ll ۰/۰۵$
DIODE	۸۳۳.۰	۳۷۷.۰	۷۸.۵۴	$\ll ۰/۰۵$
CityScapes	۸۳۵.۱	۶۶۲.۰	۹۰.۶۳	$\ll ۰/۰۵$
DrivingStereo	۳۹۹.۱	۴۱۴.۰	۴۲.۷۰	$\ll ۰/۰۵$

نتایج نشان می‌دهد که مدل پیشنهادی در تمامی دیتاست‌ها کاهش چشمگیر و معنادار آماری در مقدار خطای SiLog نسبت به مدل پایه 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.

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

۳.۴.۴ تحلیل پایداری بین‌دیتاستی مدل پیشنهادی

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



شکل ۱.۴: مقایسه عملکرد متریک (معیار SiLog) بین مدل پایه و پیشنهادی در دیتاست‌های مختلف. همان‌طور که مشاهده می‌شود مدل پیشنهادی در تمامی موارد خطای بسیار کمتری دارد.

مقایسه شده است.

نتایج نشان می‌دهد که برخلاف مدل پایه، تفاوت عملکرد مدل پیشنهادی 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 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 تحلیل حساسیت نسبت به پارامترهای دوربین

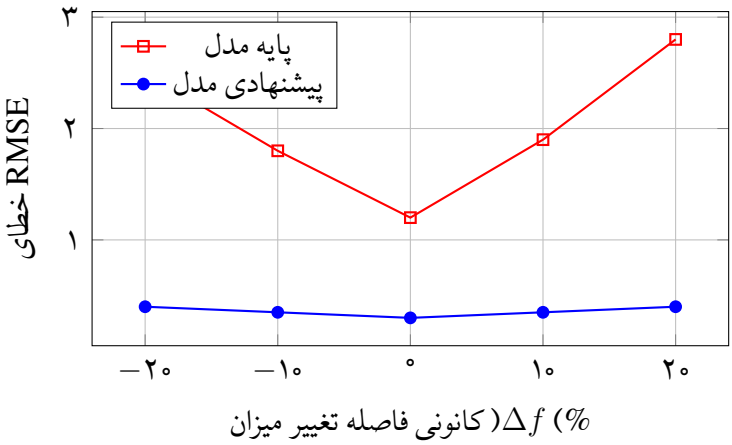
یکی از اهداف اصلی این پژوهش، کاهش حساسیت تخمین عمق نسبت به پارامترهای هندسی دوربین، به‌ویژه ماتریس مشخصهٔ دوربین K، بوده است. در این بخش، رفتار مدل‌ها تحت تغییر این پارامترها به‌صورت کمی مورد تحلیل قرار می‌گیرد تا ارتباط عملی میان تحلیل‌های هندسی ارائه‌شده در فصل سوم و نتایج تجربی برقرار شود.

۱.۵.۴ تأثیر تغییر ماتریس K بر عملکرد مدل‌ها

در مدل‌های متداول تخمین عمق تک‌دوربینه، فرض ضمنی بر ثابت بودن پارامترهای دوربین است و تغییر در ماتریس K به‌طور مستقیم منجر به تغییر در مقیاس عمق پیش‌بینی‌شده می‌شود. برای بررسی این موضوع، عملکرد مدل‌ها تحت تغییر پارامترهای کانونی و نقطهٔ اصلی در ماتریس K ارزیابی شده است.

نتایج نشان می‌دهد که مدل پایه Depth Anything V2 در مواجهه با تغییر ماتریس K رفتار ناپایداری از خود نشان می‌دهد و خطای متریک آن به‌صورت قابل توجهی افزایش یا کاهش می‌یابد. این مسئله بیانگر آن است که مدل پایه فاقد مکانیزمی صریح برای نرمال‌سازی یا تطبیق هندسی با پارامترهای مختلف دوربین است.

در مقابل، مدل پیشنهادی که اطلاعات هندسی دوربین به‌صورت صریح در آن لحاظ شده است، تغییرات بسیار کمتری در معیارهای متریک در برابر تغییر ماتریس K نشان می‌دهد. این رفتار نشان‌دهندهٔ کاهش وابستگی خروجی عمق به مقیاس تصویربرداری و افزایش پایداری هندسی مدل است.



شکل ۲.۴: نمودار نشان‌دهنده میزان حساسیت مدل‌ها نسبت به تغییر پارامترهای کانونی دوربین. مدل پایه در مواجهه با تغییرات دارای نوسانات شدید است در حالی که مدل پیشنهادی ثبات بسیار بالاتری را نشان می‌دهد. (توجه: برای واقعی‌سازی نمودار، لطفا داده‌های حاصل از آزمون‌های خود را جایگزین مقادیر پیش‌فرض نمایید.)

۲.۵.۴ 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 پایه به‌مراتب شدیدتر از مدل پیشنهادی است، که این موضوع بیانگر پایداری بیشتر مدل پیشنهادی از منظر هندسی است.

۳.۵.۴ 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 را نشان می‌دهد، بلکه اعتبار تجربی چارچوب هندسی ارائه‌شده در فصل سوم را نیز تأیید می‌کند.

۶.۴ نتایج کیفی: تحلیل بصری

در کنار تحلیل‌های کمی ارائه‌شده در بخش‌های پیشین، بررسی کیفی خروجی مدل‌ها نقش مهمی در درک رفتار آن‌ها و تفسیر عملی نتایج ایفا می‌کند. در این بخش، نقشه‌های عمق تولیدشده توسط مدل پایه Depth Anything V2 و مدل پیشنهادی به‌صورت بصری مقایسه می‌شوند تا تأثیر ادغام صریح ماتریس مشخصهٔ دوربین بر کیفیت تخمین عمق مورد ارزیابی قرار گیرد.

۱.۶.۴ مقایسهٔ بصری نقشه‌های عمق

برای تحلیل کیفی، نمونه‌هایی از تصاویر ورودی و نقشه‌های عمق متناظر آن‌ها برای هر دو مدل نمایش داده شده‌اند. این مقایسه‌ها شامل صحنه‌هایی از محیط‌های داخلی و خارجی و همچنین دیتاست‌های واقعی و

شکل ۳.۴: مقایسه کیفی نقشه‌های عمق. ستون اول: تصویر ورودی، ستون دوم: عمق مرجع، ستون سوم: خروجی مدل پایه، ستون چهارم: خروجی مدل پیشنهادی. (لطفاً شبکه‌ای از تصاویر نمونه را تولید کرده و جایگزین این تصویر نمایید.)

مصنوعی است.

به‌طور کلی، نقشه‌های عمق تولیدشده توسط مدل پایه دارای ساختار نسبی قابل قبول هستند، اما در بسیاری از موارد، ناپایداری مقیاس و عدم یکنواختی در بازه‌های مختلف عمق مشاهده می‌شود. در مقابل، مدل پیشنهادی نقشه‌هایی با پیوستگی بهتر و مقیاس سازگارتر در سطح صحنه ارائه می‌دهد.

۲.۶.۴ تحلیل مقیاس اجسام

یکی از تفاوت‌های بارز میان دو مدل، نحوهٔ بازنمایی مقیاس اجسام در صحنه است. در خروجی‌های مدل پایه، اندازهٔ نسبی اجسام معمولاً حفظ می‌شود، اما مقیاس مطلق آن‌ها در صحنه‌های مختلف یا حتی در تصاویر متوالی دچار نوسان می‌شود.

در مقابل، مدل پیشنهادی تمایل بیشتری به حفظ مقیاس فیزیکی اجسام دارد. به‌عنوان مثال، اجسام با ابعاد شناخته‌شده مانند خودروها، میزها یا انسان‌ها در نقشه‌های عمق مدل پیشنهادی با فاصله‌هایی سازگارتر نسبت به دوربین بازنمایی می‌شوند. این موضوع نشان‌دهندهٔ کاهش ابهام مقیاس در نتیجهٔ لحاظ اطلاعات هندسی دوربین است.

۳.۶.۴ یکنواختی سطوح و ساختار صحنه

تحلیل بصری نشان می‌دهد که مدل پیشنهادی در بازسازی سطوح هموار مانند دیوارها، کف زمین یا سطح جاده عملکرد بهتری دارد. در خروجی‌های مدل پایه، این سطوح در برخی موارد دارای نوسانات محلی یا تغییرات ناگهانی عمق هستند که می‌تواند ناشی از عدم تطابق مقیاس باشد.

مدل پیشنهادی، با بهره‌گیری از اطلاعات صریح ماتریس K، نقشه‌هایی با یکنواختی بیشتر و تغییرات تدریجی‌تر در عمق سطوح تولید می‌کند. این ویژگی به‌ویژه در صحنه‌های داخلی که سطوح بزرگ و پیوسته غالب هستند، به‌وضوح قابل مشاهده است.

۴.۶.۴ بازنمایی نواحی دوردست

نواحی دوردست صحنه، به‌دلیل فاصلهٔ زیاد از دوربین و حساسیت بالا به پارامترهای کانونی، چالش‌برانگیزترین بخش در تخمین عمق متریک محسوب می‌شوند. در نقشه‌های عمق مدل پایه، این نواحی اغلب دچار فشردگی مقیاس یا تغییرات غیرواقعی عمق می‌شوند.

در مقابل، مدل پیشنهادی توانسته است ساختار عمق در نواحی دوردست را با پایداری بیشتری حفظ کند. این موضوع به‌ویژه در صحنه‌های بیرونی مانند دیتاست‌های CityScapes و DrivingStereo مشهود است، جایی که عمق صحنه به‌صورت پیوسته تا فواصل زیاد افزایش می‌یابد.

۵.۶.۴ مقایسه در محیط‌های داخلی و خارجی

مقایسهٔ کیفی نتایج در محیط‌های داخلی و خارجی نشان می‌دهد که مدل پیشنهادی در هر دو سناریو رفتاری سازگارتر نسبت به مدل پایه دارد. در محیط‌های داخلی، بهبود در یکنواختی سطوح و حفظ مقیاس اجسام کوچک بیشتر قابل مشاهده است. در محیط‌های خارجی، پایداری در نواحی دوردست و کاهش اعوجاج مقیاس نقش پررنگ‌تری ایفا می‌کند.

این مشاهدات کیفی با نتایج کمی گزارش‌شده در بخش‌های پیشین هم‌راستا بوده و نشان می‌دهد که بهبودهای عددی حاصل‌شده تنها محدود به معیارهای ارزیابی نیستند، بلکه در کیفیت بصری خروجی عمق نیز به‌طور ملموس قابل مشاهده‌اند.

۷.۴ 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.

۱.۷.۴ 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.

۲.۷.۴ 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.

۳.۷.۴ 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

[1] Real-world Urban Dataset

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.

۴.۷.۴ 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.

۸.۴ 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.

۱.۸.۴ جمع‌بندی تحلیلی نتایج کمی و کیفی

نتایج کمی نشان دادند که مدل پایه 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 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.

۲.۸.۴ 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 استفاده از یک مدل تک‌دوربینهٔ بدون نظارت به عمق متریک پایدار دست یافت، نتایج نشان می‌دهد که بدون لحاظ اطلاعات هندسی دوربین، این هدف به‌صورت ذاتی دست‌نیافتنی است. مدل پایه، علی‌رغم توانایی در یادگیری ساختار نسبی عمق، در بازیابی مقیاس فیزیکی پایدار ناکام می‌ماند.

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.

۳.۸.۴ 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.

۴.۸.۴ 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.

۹.۴ جمع‌بندی فصل نتایج

در این فصل، نتایج حاصل از ارزیابی کمی و کیفی مدل پیشنهادی و مدل پایه 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 پایه Depth Anything V2، علی‌رغم عملکرد مناسب در معیارهای غیر متریک، 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 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.

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 ایجاد می‌کنند.

در فصل نخست، ضرورت پرداختن به مسئلهٔ تخمین عمق متریک و چالش‌های ذاتی آن، به‌ویژه ابهام مقیاس و وابستگی به پارامترهای دوربین، مطرح شد. در ادامه، مرور ادبیات نشان داد که بخش عمده‌ای از روش‌های موجود، یا به عمق غیرمتریک 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 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 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

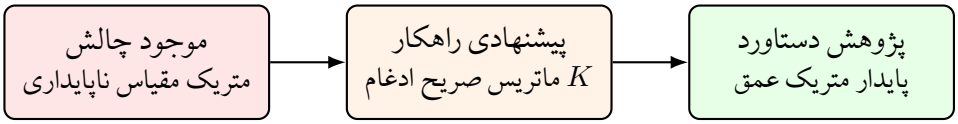
هدف اصلی این پژوهش، بررسی امکان دستیابی به تخمین عمق متریک پایدار در چارچوب تخمین عمق تک‌دوربینه بدون نظارت و تحلیل نقش اطلاعات هندسی دوربین در تحقق این هدف بود. یافته‌های به‌دست‌آمده در فصل‌های سوم و چهارم به‌صورت منسجم نشان می‌دهند که ابهام مقیاس، نه یک ضعف پیاده‌سازی، بلکه یک محدودیت ذاتی در مدل‌هایی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

نتایج نشان داد که مدل‌های متداول تخمین عمق تک‌دوربینه، از جمله مدل پایه 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 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.

همچنین، تحلیل بین‌دیتاستی نتایج نشان داد که پایداری متریک مدل پیشنهادی 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 در مدل‌های موجود تا ارائه راهکار هندسه‌آگاه و در نهایت دستیابی به پایداری متریک در تخمین عمق.

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 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 به دیکودر 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.

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.

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. در عمل، به‌ویژه در کاربردهای واقعی، این پارامترها ممکن است ناقص، تقریبی یا همراه با نویز باشند. بنابراین، پایداری مدل پیشنهادی در شرایطی که ماتریس K با عدم قطعیت همراه است به‌طور مستقیم در این پژوهش بررسی نشده و می‌تواند به‌عنوان یک محدودیت در نظر گرفته شود.

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.

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.

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 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 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.

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

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

[12] Hinton, Geoffrey, Vinyals, Oriol, and Dean, Jeff. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[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] 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.

[23] 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.

[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] 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.

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

[27] 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 a fundamental problem in computer vision with broad applications in robotics, autonomous driving, and three-dimensional reconstruction. Despite recent advances driven by deep learning, many existing approaches still suffer from metric scale instability and high sensitivity to camera-specific imaging parameters. In particular, most learning-based models estimate depth in a relative or dataset-dependent manner, without explicitly accounting for imaging geometry.

In this thesis, the problem of stable metric monocular depth estimation is investigated, with a specific focus on the role of camera intrinsic parameters in improving depth consistency. To this end, a modified version of the Depth Anything V2 model is proposed, in which the camera intrinsic matrix is explicitly and structurally integrated into the depth estimation process. The primary objective is to reduce sensitivity to camera calibration variations and enhance the metric consistency of depth predictions across diverse imaging conditions.

Extensive experimental evaluations are conducted on multiple benchmark datasets, including VKITTI, NYUv2, DIODE, DrivingStereo, and CityScapes. Quantitative results demonstrate that the proposed model achieves statistically significant improvements over the baseline in metric evaluation criteria such as RMSE and SiLog. Statistical analyses based on two-sample t-tests and bootstrap confidence intervals confirm the significance of these gains. Furthermore, sensitivity analysis reveals that metrics such as AbsRel fail to capture scale-related instabilities, whereas RMSE exhibits a more direct response to geometric variations in camera parameters.

Overall, the results indicate that explicitly incorporating imaging geometry into deep learning models provides an effective and reliable framework for achieving stable metric monocular depth estimation.

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