مقاله پژوهشی

چالش‌های موقعیت‌یابی متن فارسی در تصاویر طبیعی و اهمیت مجموعه دادگان جدید برای ارزیابی مدل‌های یادگیری عمیق

زبیر رئیسی، رسول دامنی، اسماعیل سارانی و ولی محمد نظرزهی حاد

چکیده: به دلیل پیچیدگی‌های ساختاری خط فارسی و کمبود مجموعه‌داده‌های (دادگان‌های) استاندارد و معتبر، موقعیت‌یابی متن فارسی و جداسازی کلمات در تصاویر ثبت‌شده با دوربین‌های معمولی، همچنان به عنوان یک چالش کلیدی در حوزه پردازش تصویر مطرح است. در این مقاله، ابتدا یک مجموعه دادگان جامع برای موقعیت یابی متن فارسی با نام FATD معرفی شده است.این مجموعه شامل بیش از ۲۰۰۰ تصویر متنوع است که متن‌هایی با فونت‌ها، اندازه‌ها و زاویه‌های مختلف، در شرایط محیطی متفاوت و با سطوح پیچیدگی بالا را در بر میگیرد. سپس، در مجموع شش مدل یادگیری عمیق شامل دو مدل مبتنی بر شبکه عصبی کانولوشنی (YOLOv۸ و CRAFT)، دو مدل ترانسفورمری (RRDETR و RRBDETR) و همچنین دو مدل زبان-بینایی (Qwen۲/۵-VL و Florence-۲)، تحت شرایط یکسان بر روی مجموعه دادگان معرفی شده، ارزیابی و مقایسه می شوند. نتایج ارزیابی نشان میدهد که ترانسفورمرها به قیمت هزینه محاسباتی بالاتر، عملکرد بهتر و دقیق‌تری را ارائه می‌دهند و بر بر اساس معیار ارزیابی H-mean دقتی تا ٦٥ درصد را کسب می‌کنند. در مقابل، شبکه های عصبی کانولوشنی (CNN) با سرعت پردازش مناسب، دقت رقابتی ارائه می‌کنند. همچنین علیرغم آموزش محدود مدل های جامع زبان-بینایی روی داده‌های متنی فارسی، بر اساس معیار ارزیابی H-mean این مدل‌ها در موقعیت یابی عملکرد قابل قبولی را به نمایش می‌گذارند.

***کلیدواژه:* مجموعه داده متن فارسی، موقعیت‌یابی متن در تصاویر، مدل‌های یادگیری عمیق ، مجموعه داده FATD.**

۱- مقدمه

متن به عنوان یکی از مهمترین منابع اطلاعات بصری، نقشی اساسی در تعامل انسان‌ها با محیط پیرامون ایفا می‌کند. اسناد رسمی، انواع تابلوهای راهنمایی و رانندگی، تابلوهای شهری، بیلبوردهای تبلیغاتی و ... همگی برای انتقال پیام و مفاهیم از متن و نوشته بهره می برند. از طرفی، حجم عظیم داده‌های بصری و لزوم موقعیت‌یابی، استخراج و پردازش سریع اطلاعات متنی موجود در تصاویر، استفاده از تکنیک‌های پردازش تصویر و بینایی ماشین را ناگزیر ساخته‌است و این امر خود یکی از چالش‌های اصلی این حوزه محسوب می‌شود [۱] تا [۶]. فرایند استخراج متن از تصاویر با استفاده از پردازش تصویر، در دو مرحله صورت می‌گیرد که عبارتند از : اول موقعیت‌یابی دقیق محدوده‌های حاوی نوشته در تصویر و دومبازشناسی و استخراج کلمات و حروف و تبدیل آنها به رشته‌های متنی قابل خواندن. فرایند موقعیت‌یابی و شناسایی متن در تصاویر به‌دلیل تنوع رنگ، فونت، اندازه، زاویه نگارش و پیچیدگی پس‌زمینه، با چالش‌های فنی پیچیده‌ای مواجه است.

عملکرد روش‌های سنتی مبتنی بر یادگیری ماشین، در مواجهه با پیچیدگی‌های موجود در تصاویر واقعی، اغلب با محدودیت‌هایی مواجه است [۷] و [۸]. اما پیشرفت‌های اخیر در حوزه یادگیری عمیق، بهبود قابل توجهی در عملکرد سیستم‌های بازشناسی متن و بسیاری از کاربردهای دیگر، به‌ویژه در شرایط پیچیده و متنوع ایجاد کرده است [۵] تا [۱۶]. با وجود این، عمده پژوهش‌ها در این زمینه بر روی زبان‌هایی با خط لاتین متمرکز بوده و مجموعه داده‌های گسترده‌ای برای ارزیابی ای نمدل‌ها در این زبان‌ها توسعه یافته است. در مقابل، زبان‌های غیر لاتین با خطوط راست به چپ[1] (RTL) و حروف متصل مانندفارسی، عربی و اردو، دارای ویژگی‌های منحصر به فرد و پیچیدگی‌های ساختاری خاصی هستند که بازشناسی متن در آن‌ها را به مسئله‌ای چالش‌برانگیزتر تبدیل می‌کند. حروف متصل به هم در موقعیت‌های مختلف، علائم نگارشی متنوع، تغییرات شکل حروف در کلمات مختلف و هم‌پوشانی فضایی حروف از جمله این پیچیدگی‌ها در بازشناسی متن فارسی هستند (شکل ۱).

این دشواری‌ها، ضرورت توسعه روش‌های اختصاصی یا معماری یادگیری عمیق متناسب برای موقعیت یابی و بازشناسی متن در زبان‌های غیر لاتین از جمله زبان فارسی را بیش از پیش آشکار می‌سازد.

ساختار خط فارسی تفاوت‌های اساسی با خطوط لاتین ندارد. وجود نقاط با تعداد و موقعیت‌های مختلف درحروف، جهت‌گیری متنوع حروف ووجود اتصال در حروف افقی خط فارسی، از جمله تمایزهای کلیدی این دو ساختار نوشتاری هستند. به دلیل این ویژگی‌های منحصر به فرد، به کارگیری الگوریتم‌های بازشناسی متن توسعه یافته برای خطوط لاتین، در تشخیص خودکار متن فارسی ناکارآمد بوده و با چالش جدی مواجه است. در نتیجه، دقت این مدل‌ها بدون آموزش جامع بر روی مجموعه دادگان فارسی و تنظیم دقیق پارامترها ممکن است به صورت چشمگیری کاهش بیابد.

همان‌گونه که در شکل ۱ نشان داده شده است، خط فارسی دارای پیچیدگی‌های ساختاری متعددی است. اتصال چند جانبه حروف (اتصال

این مقاله در تاریخ ۲۳ مرداد ماه ۱۴۰۴ دریافت و در تاریخ ۸ آبان ماه ۱۴۰۴ بازنگری شد.

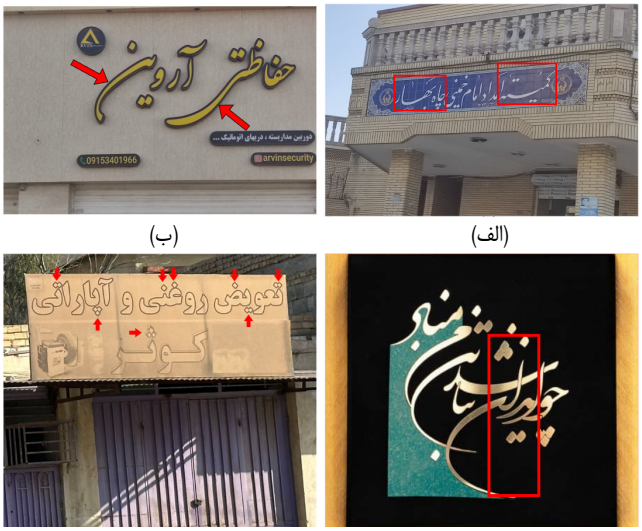
زبیر رئیسی (نویسنده مسئول)، دانشکده مهندسی دریا، دانشگاه دریانوردی وعلوم دریایی چابهار، چابهار، ایران، (email: zobeir.raisi@cmu.ac.ir).

رسول دامنی، دانشکده مهندسی دریا، دانشگاه دریانوردی وعلوم دریایی چابهار، چابهار، ایران، (email: damani@cmu.ac.ir).

اسماعیل سارانی، دانشکده مهندسی دریا، دانشگاه دریانوردی وعلوم دریایی چابهار، چابهار، ایران، (email: sarani@cmu.ac.ir).

ولی محمد نظرزهی حاد، دانشکده مهندسی دریا، دانشگاه دریانوردی وعلوم دریایی چابهار، چابهار، ایران، (email: v.nazarehi@gmail.com).

1. Right to Left



(الف) (ب) (ج) (د)

شکل ۱: برخی از پیچیدگی‌های منحصر به فرد و ساختاری خط فارسی که بهترین مدل‌هاییادگیری عمیق در موقعیتیابی کلمات را به به چالش می‌کشد. این چالش‌ها عبارتند از: (الف)اتصال معنایی دو کلمه به ظاهر مجزا در قالب یک کلمه مانند "چاه بهار"،(ب) بیرون زدگی برخیحروف از خط مبنای افقی مانند حروف "ی" و "ن"، (ج) تداخل و همپوشانی فضایی حروف وکلمات در بعضی از نوشته‌های فارسی، و (د) تعدد حروف دارای نقطه و گاهی مشابه که تعداد و محل نقاط تنها وجه تمایز آنها بوده و تشخیص آنها از هم با چالش همراه است، مانند حروف نشان‌داده شده که با پیکان قرمز.

به حرف پیشین، حرف پسین یا هر دو)، تنوع شکل حروف در موقعیت‌های مختلف در یک کلمه، هم‌پوشانی فضایی حروف و کلمات از جمله این پیچیدگی‌هاست. این ویژگی‌ها، به ویژه مایز برخی از حروف مشابه صرفاً بر اساس تعداد و محل قرارگیری نقاط، تفاوت‌های چشمگیری را بین خط فارسی و خطوط لاتین ایجاد و کارایی الگوریتم‌های پیشرفته تشخیص متن لاتین برای تشخیص متن فارسی را با چالش جدی مواجه می‌کند.

با وجود شباهت‌های ظاهری بین خطوط فارسی، عربی و اردو، تفاوت‌های ساختاری در شکل حروف و قواعد نگارشی این خطوط وجود دارد و این موضوع، استفاده از مجموعه دادگان یک زبان برای آموزش مدل‌های تشخیص متن زبان‌های دیگررا با ناکارآمدی نسبی چالش مواجه خواهد نمود. لذا، کارآمدی استفاده از مجموعه دادگان خطوط عربی و اردو، برای بازشناسی متون فارسی در تصاویر واقعی محدود است. از طرفی، مجموعه داده‌های استاندارد و عمومی برای بازشناسی متن فارسی نیز محدودند. پژوهش‌های پیشین در حوزه مجموعه دادگان فارسی عمدتاً بر تولید مصنوعی داده‌ها [۱۷]، تمرکزصرف بر وظایف موقعیت‌یابی متن در تصاویر خاص [۱۸] یا بازشناسی انواع خاصی از متن مانند اسناد [۱۹] و [۲۰] متمرکز بوده‌اند.

در این مقاله، یک مجموعه داده جدید با عنوان FATD [1] معرفی می‌شود که برای موقعیت‌یابی و بازشناسی نوشته‌های متنوع فارسی در تصاویر، از منابع و محیط‌های گوناگون تهیه شده است.

این مجموعه داده شامل طیف گسترده‌ای از تصاویر متن فارسی با کیفیت‌های متفاوت است که از محیط‌های داخلی مانند تصاویر محصولات فروشگاهی و راهروها و محیط‌های خارجی شامل تابلوهای سردر فروشگاه‌ها، تابلوهای راهنمایی و رانندگی، تابلوهای شهرداری و دیوارنوشته‌ها جمع‌آوری شده‌اند. تصاویر مذکور با استفاده از دوربین موبایل، هم در حالت ایستا و هم درحال حرکت ثبت شده‌اند.

این تصاویر، تنوع بالایی در سطوح نویز، روشنایی و زاویه چرخش داشته و طیف وسیعی از چالش‌های خط فارسی را در بر می‌گیرند. از این رو، این مجموعه دادگان برای آموزش و ارزیابی مدل‌های موقعیت‌یابی و بازشناسی متن فارسی در شرایط واقعی، بسیارحائز اهمیت است و امکان توسعه مدل‌های دقیق‌تر و کارآمدتر را فراهم می‌آورد و به این ترتیب، به کاربردهای عملی این فناوری در حوزه‌های مختلف کمک شایانی می‌نماید. به طور خلاصه، نوآوری‌های اصلی این مقاله به شرح زیر است:

- معرفی و عرضه عمومی مجموعه دادگان متنوع و جامع FATD با پوشش انواع چالشهای خط و متون فارسی که علاوه بر ارزیابی مدل‌های متنوع موقعیت یابی و بازشناسی متون موضوع این مقاله، در آینده نیز مورد استفاده پژوهشگران در تحقیقات مرتبط خواهد بود. FATD اولین مجموعه دادگان جامع و در دسترس عموم است که به‌طورخاص برای موقعیت‌یابی متن فارسی طراحی شده و چالش‌های موجوددراستفاده از مجموعه‌های داده لاتین را نیز در نظر گرفته است.
- این مقاله، با مرور پیشرفت‌های گذشته و اخیر در زمینه موقعیت‌یابی متن در تصاویر، هم برای خطوط لاتین و هم خطوط غیر لاتین، درک جامعی از وضعیت موجود در این حوزه را به خواننده ارائه می‌دهد.
- شش مدل پیشرفته موقعیت‌یابی متن در سه دسته‌ی CNN[2]ها (CRAFT و YOLOv۸)، ترنسفورمرها (RRDETR و RRBDETR) و مدل‌های زبان-بینایی (Qwen۲/۵-VL و Florence)، بر روی مجموعه داده معرفی شده FATD به صورت کمی و کیفی ارزیابی و مقایسه می‌شوند.

بر اساس نتایج حاصل از ارزیابی‌ها، برای بهبود عملکرد سیستم‌های موقعیت‌یابی متن فارسی، مسیرهایی برای تحقیقات آینده به پژوهشگران این حوزه، پیشنهاد شده است.

۲- پیشینه پژوهش

۲-۱ موقعیت‌یابی متن درتصاویر

روش‌های موقعیت‌یابی متن در تصاویر را می‌توان به دو دسته کلی تقسیم کرد: روش‌های یادگیری کلاسیک و روش‌های یادگیری عمیق [۷]، [۸] و [۲۱]. روش‌های کلاسیک معمولاً از دو رویکرد اصلی پنجره لغزنده (SW)[3] و مولفه متصل (CC)[4] استفاده می‌کنند. در رویکرد پنجره لغزنده، تصویر به صورت سیستماتیک با یک پنجره اسکن می‌شود تا نواحی محتمل حاوی متن شناسایی شوند. سپس، ویژگی‌های محلی مانند هیستوگرام گرادیان جهت‌یافته[5] (HOG) استخراج و به یک طبقه‌بندی‌کننده مانند ماشین بردار پشتیبان[6] (SVM) ارائه می‌شود. در رویکرد مولفه متصل، تصویر به نواحی کوچکتری تقسیم‌شده و نواحی با ویژگی‌های بصری مشابه گروه‌بندی می‌شوند. این گروه‌ها سپس به عنوان نامزدهای متن در نظر گرفته شده و با استفاده از طبقه‌بندی‌کننده‌ها، اعتبارسنجی می‌شوند. روش‌های مبتنی بر مولفه متصل با گروه‌بندی نواحی تصویر بر اساس ویژگی‌های مشترک، روشی کارآمد برای موقعیت‌یابی متن در تصاویر ارائه می‌دهند. این روش‌ها به طور گسترده در

1. Farsi Text Detection
2. Convolution Neural Network
3. Sliding Window
4. Connected Components
5. Histogram Oriented Gradients
6. Support Vector Machines

بسیاری از کارهای موقعیت‌یابی متن مورد استفاده قرار می‌گیرند؛ روش‌هایی مثل SWT [۲۱] و MSER [۲۲] از جمله نمایندگان اصلی این دسته هستند. با این وجود، این روش‌ها نیز با چالش‌هایی مواجه هستند. به عنوان مثال، آنها ممکن است در موقعیت‌یابی متن با جهت‌های دلخواه دچار مشکل شده و یا تشخیص‌های نادرستی ایجاد کنند [۴].

پیشرفت‌های اخیر در حوزه یادگیری عمیق، تحولات چشمگیر یرادر زمینه موقعیت‌یابی متن در تصاویر ایجاد نموده‌اند [۵] و[۶]. این روش‌ها نسبت به الگوریتم‌های سنتی یادگیری ماشین، مزایایی همچون معماری ساده‌تر، توانایی موقعیت‌یابی متن در زوایای مختلف و عملکرد بهبودیافته بر روی مجموعه دادگان مصنوعی را ارائه می‌دهند. به طورکلی، روش‌های موقعیت‌یابی متن مبتنی بر یادگیری عمیق را می‌توان به دو دسته اصلی تقسیم‌بندی نمود: روش‌های رگرسیون کادر مرزی[۱] و روش‌های مبتنی بر بخش‌بندی[۲]. اغلب این روش‌ها، معماری خود را از الگوریتم‌های عمومی موقعیت‌یابی اشیاء مانند Faster-RCNN [۲۳]، YOLO[۲۴]، SSD [۲۵]، شبکه‌های کاملاً کانولوشنی (FCN) [۲۶]، Mask-RCNN [۲۷] و روش‌های موقعیت‌یابی مبتنی بر ترانسفورماتورها [۲۸] و [۲۹] الهام گرفته‌اند.

روش‌های رگرسیون کادر مرزی [۹]، [۱۰] و [۳۰]، متن را به عنوان شیء در نظر گرفته و کادرهای محصورکننده‌ی احتمالی را به‌طور مستقیم پیش‌بینی می‌کنند. این روش‌ها هر چند کارآمد هستند، اما ممکن است در مواجهه با نمونه‌های متن با جهت‌های دلخواه، با مشکلاتی مواجه شوند. برخی از معماری‌های یادگیری عمیق موفق به رفع این محدودیت‌ها شده‌اند، به‌طوری که روش‌هایی مانند EAST [۹] و TextBoxes++ [۱۰] در مجموعه داده‌های مختلف با نمونه‌های متن جهت‌دار عملکرد خوبی از خود نشان داده‌اند.

در سال‌های اخیر، برای موقعیت‌یابی متن، مدل‌های پیشرفته‌ای [۱۲] [۱۳] با استفاده از آشکارسازهای مبتنی بر ترانسفورمر مانند DETR [۲۸] و Deformable-DETR [۲۹] توسعه یافته‌اند که قادرند متن را با اشکال دلخواه موقعیت‌یابی کنند، بدون آنکه به طراحی معماری پیچیده‌ای نیاز داشته باشند. روش‌های مبتنی بر تقسیم‌بندی معنایی تصویر به نواحی مختلف [۳۱] و [۳۲] نیز از رویکردهای محبوب در موقعیت‌یابی متن هستند. در این روش‌ها، هر پیکسل از تصویر به یک کلاس (مثل متن یا پس‌زمینه) اختصاص داده می‌شود. این رویکرد به ویژه برای موقعیت‌یابی متن در تصاویر پیچیده و با اشکال نامنظم بسیار مناسب است. با این حال، به دلیل اتصال احتمالی بین حروف، این روش‌ها ممکن است در جداکردن کلمات مجاور با مشکل مواجه شوند. بسیاری از این روش‌ها از شبکه‌های عصبی کانولوشنی کاملاً متصل مانند Mask R-CNN [۲۷] و U-Net [۳۳] الهام گرفته‌اند که پیش‌تر در زمینه قطعه‌بندی اشیاء به کار رفته‌اند.

پیشرفت‌های اخیر در مدل‌های بینایی-زبان[۳](VLM) منجر به توسعه مدل‌های یکپارچه و قدرتمندی برای تشخیص اشیا، مکان‌یابی بصری و شناسایی متن در تصاویر شده است. مدل‌های ترنسفورمرمحور مانند Florence [۳۴] و نسخه پیشرفته‌تر آن Florence-۲ [۳۵] از معماری مبتنی بر «دستورمحور[۴]» (پرامپت) و « زنجیروار[۵]» (Seq2Seq) بهره می‌برند که با مجموعه‌داده‌های بسیار بزرگ و غنی از حاشیه‌نویسی (برچسب‌گذاری) آموزش دیده‌اند. این مدل‌ها قادرند وظایفی همچون مکان‌یابی اشیا، توصیف تصویر، مکان‌یابی ناحیه‌ای، بخش‌بندی و OCR[۶] OCR را با دقت بالا چه در حالت بدون آموزش مجدد[۷] و چه در حالت تنظیم دقیق[۸] انجام دهند [۳۵].

به‌ویژه Florence-۲ از طریق نمایشی کپارچه و مکانیزم دستوردهی مبتنی بر پرامپت، توانایی چشمگیری در انجام طیف گسترده‌ای از وظایف از جمله تشخیص اشیا، تولید توضیح برای نواحی متراکم، OCR و بخش‌بندی ناحیه‌ای از خود نشان داده است [۳۵].

به‌طور مشابه، Qwen۲/۵-VL [۳۶] با گسترش قابلیت‌های مجموعه مدلQwen-VL [۳۷]، بهبودهای قابل توجهی در زمینه‌های پردازش اسناد، مکان‌یابی اشیا، OCR چندزبانه و درک ویدئو ارائه داده است. این مدل با پشتیبانی از تشخیص دقیق اشیا و مکان‌یابی بصری، قادر به انجام وظایفی همچون تشخیص متن درتصاویر و استخراج پیچیده اطلاعات از اسناد با چینش‌ها و جهت‌گیری‌های مختلف است. در ارزیابی‌های انجام‌شده، مدل‌های Qwen۲/۵-VL (به‌ویژه نسخه‌هایB۳۲ و B۷۲) در وظایف استخراج ساختاریافته مبتنی بر OCR به دقتی در حدود ۷۵ درصد دست یافته‌اند؛ عملکردی که هم‌تراز با GPT-۴o [۳۸] است و نشان از توان رقابتی این مدل‌ها در تشخیص متن صحنه و درک اسناد دارد [۳۷]. علاوه بر این، Qwen۲/۵VL در زمینه‌ی تشخیص اشیاء به‌صورت بدون آموزش مجدد و در استدلال مکانی و موقعیت‌یابی، عملکردی برتر نسبت به سایر مدل های مشابه از خود نشان می‌دهد . در این مقاله جهت ارزیابی، دو مدل کانولوشنی شامل (CRAFT و YOLOv۸)، دومدل ترنسفورمری (RRDETR و RRBDETR) و دومدل زبان-بینایی (Qwen۲/۵VL و Florence) انتخاب شده‌اند که برخی از ویژگی های آنها شامل خلاصه معماری، مزایا و معایب هر کدام در جدول ۱ آورده شده است.

۲-۲ مروری بر مجموعه دادگان موقعیت‌یابی و بازشناسی متن با رسم الخط غیر از فارسی

۲-۲-۱ مجموعه دادگان متونلاتین

برای سنجش عملکرد سیستم‌های موقعیت‌یابی متن لاتین، مجموعه داده‌های متعددی به عنوان معیار ارزیابی دردسترس است. از جمله این مجموعه‌ها می‌توان به ICDAR۱۳ [۴۲] برای موقعیت‌یابی متن افقی با برچسب‌گذاری (حاشیه‌نویسی) کادرهای مرزی مستطیلی، ICDAR۱۵ [۴۳] برای موقعیت‌یابی تصادفی متن در تصاویر با حاشیه‌نویسی کادرهای مرزی چهارضلعی، و COCO-Text [۴۴] باحاشیه‌نویسی مستطیلی و چالش‌های پیچیده‌تر اشاره کرد. مجموعه دادگان COCO-Text معیاری مهم برای سنجش قابلیت تعمیم‌پذیری مدل‌های موقعیت‌یابی متن محسوب می‌شود. علاوه بر این، مجموعه‌های دادگان Text-OCR[۴۵] با حاشیه‌نویسی‌های چهارضلعی و چندضلعی، برای موقعیت‌یابی متن با اشکال دلخواه طراحی شده است. همچنین، مجموعه داده‌های تخصصی مانند Total-Text [۴۶] و CTW۱۵۰۰ [۴۷] به‌طور ویژه برای موقعیت‌یابی متن منحنی توسعه یافته‌اند. در نهایت، مجموعه داده HierText [۴۸] با ارائه حاشیه‌نویسی‌های سلسله مراتبی خطوط متن، به عنوان یک معیار جامع برای ارزیابی مدل‌های پیشرفته

1. Bounding Box
2. Segmentation
3. Vision Language Model
4. Prompt
5. Sequence to Sequence
6. Object Character Recognition
7. Zero-Shot Learning
8. Fine-Tuning

موقعیت‌یابی متن مورد استفاده قرار می‌گیرد [۴۹].

علاوه بر استفاده از مجموعه دادگان واقعی، محققان از مجموعه دادگان مصنوعی نیز برای بهبود عملکرد مدل‌های موقعیت‌یابی متن بهره می‌گیرند. مجموعه‌های دادگان مصنوعی SynthText [۵۰] و MJSynth [۵۱] به طور گسترده‌ای در مرحله پیش‌آموزش مدل‌های موقعیت‌یابی و شناسایی متن در تصاویر مختلف کاربرد دارند. این مجموعه‌ها با فراه مکردن حجم عظیمی از داده‌های متنوع، به افزایش دقت و تعمیم‌پذیری مدل‌ها کمک می‌نمایند.

۲-۲-۲ متون چندزبانه

برای ارزیابی عملکرد مدل‌های موقعیت‌یابی متن چندزبانه، مجموعه‌های دادگان متنوعی به کار گرفته شده‌اند. ICDAR۱۷-MLT [۵۲] و ICDAR۱۹-MLT [۵۳] دو نمونه از این مجموعه‌ها شامل زبان‌های عربی، لاتین، چینی، ژاپنی، کره‌ای، بنگالی وهندی هستند. مجموعه داده‌ی MSRATD۵۰۰ [۵۴]، که شامل خطوط متن طولانی به زبان‌های انگلیسی و چینی است، به عنوان یک معیار استاندارد برای موقعیت‌یابی متن چرخش‌یافته شناخته می‌شود. علاوه براین، مجموعه های دادگان دیگری نیز همانند [۵۳]، [۵۵] تا [۵۷] وجود دارند که عمدتاً شامل خطوط انگلیسی و چینی بوده و به صورت خاص برای موقعیت‌یابی متن طراحی شده‌اند. در میان این مجموعه دادگان، MSRATD۵۰۰ [۵۴] به دلیل دارا بودن حاشیه‌نویسی دقیق خطوط متن و کاربرد گسترده درپژوهش‌ها، به عنوان یکی از مهم‌ترین معیارهای ارزیابی در این حوزه محسوب می‌شود.

۲-۲-۳ متون راست به چپ

زبان‌های عربی، اردو و فارسی از رسم‌الخط‌های مشابه مبتنی بر الفبای عربی بهره می‌برند، اما از نظرساختارزبانی، واج‌شناسی و نحوه نگارش تفاوت‌های قابل توجهی دارند. در مقایسه با زبان‌های چپ‌به‌راست مانند لاتین و چینی، دسترسی به مجموعه داده‌های معیار عمومی و آزاد برای این زبان‌ها محدودتر است. هرچند پژوهش‌های گسترده‌ای بر روی این زبان‌ها با استفاده ازمجموعه داده‌های اختصاصی صورت گرفته است، اما فقدان داده‌های عمومی و در دسترس، همچنان یکی از محدودیت‌ها یاساسی در توسعه و ارزیابی مدل‌های مرتبط با این زبان‌ها به شمار می‌آید. در میان این زبان‌ها، منابع داده‌ای برای عربی نسبت به فارسی غنی‌تر است. به عنوان مثال، مجموعه داده‌های ICDAR۱۷-MLT و ICDAR۱۹-MLT به عنوان منابع ارزشمند برای ارزیابی مدل‌های موقعیت‌یابی متنعربی شناخته می‌شوند. مجموعه‌های ARASTEC [۵۸] و ARASTI [۵۹] که به‌ترتیب برای موقعیت‌یابی کاراکتر و کلمات عربی در تصاویر طبیعی طراحی شده‌اند، دردسترس عموم نیستند. در حوزه‌ی زبان اردو، مجموعه‌های داده IIITH [۶۰] و UPTI [۶۱] از منابع مهم پژوهشی به شمار می‌آیند که شامل نمونه‌های واقعی و مصنوع یهستند. اخیراً، رحمان و همکاران [۶۲] دو مجموعه داده‌ی جدید با عناوین UTRSet-Real و UTRSet-Synth را به صورت عمومی در دسترس پژوهشگران قرار داده‌اند.

۲-۲-٤ متن فارسی در تصاویر

با مرور پژوهش‌های پیشین در زمینه‌ی تشخیص و شناسایی متن فارسی، می‌توان مشاهده کرد که این حوزه هم‌زمان با پیشرفت‌های قابل‌توجه، با چالش‌های فنی و داده‌ای نیز روبه‌رو است. مطالعات اخیر عمدتاً بر توسعه‌ی چارچوب‌های کارآمد برای شناسایی متن فارسی در قالب‌های متنوع از جمله اسناد چاپی، تصاویر صحنه و ویدئوها متمرکز بوده‌اند.

بیشتر پژوهش‌ها در زمینه تشخیص متن فارسی به تصاویر اسنادی اختصاص دارند [۶۳] و [۶۴]. برای مثال، مجموعه داده SUT که در [۶۳] معرفی شده است، یک مجموعه داده‌ی بزرگ مقیاس برای تصاویر اسنادی فارسی است که شامل ۶۲,۴۵۳ تصویر در ۲۱ کلاس مختلف است. این مجموعه‌ها از وظایفی همچون طبقه‌بندی اسناد و OCR پشتیبانی می‌کند. با استفاده از مدل‌های Tesseract و EasyOCR، نویسندگان این مقاله به دقت بالا و نرخ خطای کاراکتر پایینی دست یافته‌اند. مقاله [۶۴] یک شبکه عصبی عمیق مبتنی بر ترانسفورمر برای شناسایی نوری کاراکترهای فارسی (OCR) ارائه می‌دهد که نسبت به روش‌های دیگران، دقت بالاتری را بر روی مجموعه داده‌های اسنادیکسب کرده است. برخی مطالعات بر موقعیت یابی متن در تصاویر و ویدیوها تمرکز دارند. مقاله [۶۵] یک چارچوب نوآورانه برای تشخیص و مکان‌یابی متن فارسی با استفاده از معماری YOLOv۵ ارائه کرده است که به کمبود مدل‌های مؤثر در روش‌های موجود پرداخته و عملکرد مدل پیشنهادی خود را بر روی یک مجموعه داده جدید از ویدیوهای خبری ارزیابی و نتایج امیدوارکننده‌ای در این حوزه نشان داده است.

پژوهش‌های اخیر، بر تشخیص متن در تصاویر گرفته‌شده از محیط های طبیعی و پرچالش‌تر نیز تمرکز داشته‌اند. به عنوان نمونه ، مجموعه داده‌ی PESTD [۱۸] یکیاز منابع ارزشمند دنیای واقعی برای موقعیت یابیو شناسایی متن به شمار می‌آید که شامل تصاویر دو زبانه فارسی–انگلیسی است. این مجموعه دادهکه از تصاویر ثبت شده در خیابان‌های تهران تشکیل شده است، شامل بیش از ۱۲,۰۰۰ تصویر در دو مجموعه بوده و به‌طور ویژه با هدف افزایش دقت شناسایی متن در تابلوهای ترافیکی طراحی شده‌است. استفاده از شبکه‌های عصبی کانولوشنی (CNN)، به ویژه الگوریتم Tiny-YOLOv۳، امکان آموزش و ارزیابی کارآمد این مجموعه داده را فراهم نموده و به نتایج عملکردی قابل توجهی در زمینه‌ی شناسایی متن فارسی دست یافته است . به هرحال، این مجموعه داده به‌طور ویژه برای موقعیت‌یابی و تشخیص علائم راهنمایی و رانندگی طراحی‌شده و محدود به تصاویر ترافیکیاست. در حالی که بیشتر تابلوها شامل متون فارسی درکنار علائم نیز می‌باشند و تشخیص آن ها حیاتی است، بنابراین بهتر است تصاویر گسترده‌تری از سیستم‌های شناسایی تابلوهای ترافیکی، از جمله محیط‌های چندزبانه و فرمت‌های متنوع تابلوها، مورد توجه قرار گیرد تا ایمنی جاده‌ها در مناطق مختلف تضمین شود [۱۸]. یکی از مقالات اخیر که به کار ما نزدیک است، در مرجع [۶۶] معرفی شده است، یک مجموعه دادگان جدید برای تشخیص متن فارسی در تصاویر گرفته شده از خیابان را معرفی می‌کند که برخی از کمبودهایمجموعه داده‌های پیشین در زمینه موقعیت‌یابی را برطرف کرده است. این مجموعه داده شامل ۱۱۸۲ تصویر حاشیه‌نویسی‌شده با کادر محصور کننده مستطیلی در فرمت الگوریتم YOLO است که پیشرفت سیستم‌های موقعیت‌یابی متن فارسی را از طریق فونت‌ها و پس‌زمینه‌های متنوع تسهیل می‌کند. کاملا متفاوت از مقالات قبلی، مجموعه‌داده‌ی ITDR-Synth [۱۷]، شامل تصاویر مصنوعی ایجاد شده با فونت‌ها و رنگ‌های مختلف، نیز به عنوان یک منبع قابل دسترسی برای تحقیقات در زمینه موقعیت‌یابی و شناسایی متن فارسی محسوب می‌شود. این مجموعه داده شامل ۶۱۰۰ تصویر برای موقعیت‌یابی و ۴۰۲۲۰ تصویر برای خواندن متن فارسی است.

به هر حال، با وجود پیشرفت‌های قابل توجه در تشخیص و شناسایی متن فارسی، این حوزه همچنان با چالش‌هایی مانند توسعه مجموعه داده

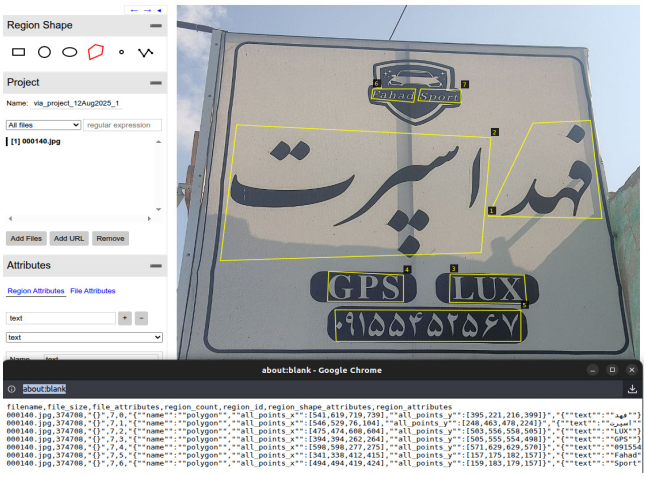
شکل ۱: نمونه‌هایی از چالش‌های گوناگون متن فارسی در تصاویر شامل: متن عمودی، متن با جهت دلخواه، وضوح پایین ، انسداد جزئی، تغییرات نور، فونت های مختلف و پیچیده، مات شدن و دست نوشته که همگی در موقعیت یابی مدل های یادگیری عمیق مشکل‌ساز هستند.

محیط‌های شهری و برون شهری از جمله تابلوهای خیابانی، تابلوهای دیواری، علائم راهنمایی و رانندگی در خیابان‌ها و جاده‌ها و تابلوهای تبلیغاتی را در بر می‌گیرند. تنوع تصاویر مورد استفاده در این مجموعه داده، در شکل ۲ نشان‌داده شده است. در این پژوهش، برای حاشیه‌نویسی (برچسب‌گذاری) تصاویر، از نرم‌افزار متن باز معرفی شده در مرجع [۶۷] استفاده شده است.

در حوزه‌ی موقعیت‌یابی متن فارسی در تصاویر، استفاده از کادر محصورکننده چهارضلعی با چهار رأس به صورت $g=[x_1,y_1,x_2,y_2,x_3,y_3,x_4,y_4]$ در فرآیند حاشیه‌نویسی محدوده‌ی متن واقعی، برای نخستین بار در این مجموعه داده به کار رفته است. شکل ۳ نمونه‌ای از حاشیه‌نویسی تصویر با استفاده از این نرم‌افزار را به همراه مختصات رئوس و متن هر کلمه نمایش می‌دهد. همانطور که در جدول ۱ نشان داده شده‌است، مجموعه داده‌ی FATD شامل ۲۰۰۴ تصویر است که ۱۵۰۴ تصویر برای آموزش و ۵۰۰ تصویر برای ارزیابی مدل‌ها در نظر گرفته شده‌اند. برای تسهیل استفاده محققان، حاشیه‌نویسی‌ها در قالب‌های مختلفی ارائه شده‌اند، از جمله فرمت ".txt" استاندارد ICDAR [۴۳]، [۵۲] و [۵۳]، برای هر تصویر، فرمت ".json" سازگار با Microsoft COCO [۶۸] و فرمت YOLO [۲۴] شامل: کادر محصورکننده مستطیلی و کادر محصورکننده مستطیلی چرخش‌یافته.

در فرآیند جمع‌آوری و آماده‌سازی مجموعه داده‌ی FATD به چالش های متعددی که ممکن است برای موقعیت‌یابی متن در تصاویر واقعی ظاهر شود، توجه شده است. این چالش‌ها شامل:انسداد، جهت‌گیری‌های مختلف، وضوح پایین، تغییرات روشنایی، فونت‌های پیچیده و وجود متن‌های چندزبانه هستند. شکل ۴ برخی از این چالش‌ها را به تصویر می‌کشد. برای تشابه حداکثری مجموعه داده‌ی معرفی شده در این مقاله با داده‌های استاندارد لاتین و ملاحظه‌ی چالش‌های آن مجموعه‌های داده، تصاویر مورد استفاده، با دوربین‌های متفاوت، در زمان‌ها و مکان‌های متفاوت و با وضوح متفاوت گرفته شده‌اند. در شکل ۵، تابع چگالی احتمال[3] (PDF) وضوح[4] و ابعاد تصاویر این مجموعه داده آورده شده است. همانطور که نشان داده شده‌است تصاویر این مجموعه داده از تنوع بالایی در وضوح و ابعاد برخوردارند. قابل ذکر است همانطور که در شکل‌های ۲ و ۴ مشاهده می‌شود و در نمودار توزیع نسبت ارتفاع به عرض شکل ۶ نشان داده شده است، ابعاد متن‌های تصاویر مجموعه داده آماده‌شده نیز دارای گستردگی متفاوتی هستند.

شکل ۲: نمونه تصاویر مورد استفاده در مجموعه داده FATD. ردیف بالا نمونه تصاویر محیط‌های داخلی شامل فروشگاه ها و پاساژها و ردیف پایین نمونه تصاویر از محیط های بیرونی با استفاده از دوربین گوشی‌های متفاوت موبایل گرفته شده‌اند. کادرهای محصور کننده دور متن با رنگ قرمز نشان داده شده‌اند.



شکل ۳: نمونه‌ای از حاشیه‌گذاری یک تصویر با استفاده از نرم افزار متن باز VIA [۶۷]. ردیف بالا کادرهای محصور کننده چهار ضلعی کلمات و ردیف پایین داده‌های حاصل از حاشیه نویسی این تصاویر را نشان می‌دهد.

تصاویر جامع یا طراحی معماری مناسب برای پوشش تمام نمونه‌های متن در تصاویر محیطی روبه‌رو است. تحقیقات بیشتری برای رفع این شکاف‌ها و افزایش کاربرد عملی این فناوری‌ها ضروری است.

۳- مجموعه داده‌ی پیشنهادی برای موقعیت‌یابی متن فارسی (FATD)

این بخش به معرفی مجموعه داده جامع FATD اختصاص دارد. این مجموعه به طور ویژه چالش‌های موقعیت‌یابی متن فارسی را مورد توجه قرار داده‌است. جزئیات بیشتر در مورد این مجموعه داده در ادامه ارائه خواهد شد. FATD شامل تصاویر متعددی از محیط‌های داخلی و خارجی و حاوی نمونه‌های متنی متنوع و چالشی است. تصاویر داخلی[1] در بر گیرنده‌ی نمونه‌های متنی متراکم و کوچک از تابلوهای فروشگاه‌ها، ویترین مغازه‌ها، و نوشته‌های روی بسته‌بندی محصولات است. در حالی که تصاویر خارجی[2] طیف گسترده‌ای از موارد چالش‌برانگیز در

1. Indoor Images
2. Outdoor Images
3. Probability Density Function
4. Resolution

جدول ۱: ویژگی‌های مدل های یادگیری عمیق مورد ارزیابی. مدل هایی که با علامت * مشخص شده‌اند به صورت محلی آموزش داده شده‌اند.

مدل	خلاصه معماری	مزایا	معایب
CRAFT [۱۱]	مبتنی بر شبکه کانولوشنی با معماری VGG16 [۳۹] و U-Net [۴۰]	دقت بالا در تشخیص متن‌های منحنی و نامنظم قابلیت تعمیم خوب روی رسم‌الخط‌های متفاوت خروجی قابل تفسیر به‌صورت نقشه حرارتی.	سرعت پایین‌تر به دلیل پیش‌بینی در سطح پیکسل نیاز به پردازش پس‌از–شبکه برای گروه‌بندی حروف. ضعف در تشخیص حروف بسیار کوچک یا متراکم.
YOLOv۸* [۲۴]	مبتنی بر معماری کانولوشنی DarkNet [۴۱]	سرعت پردازش بالامناسب برای کاربردهای بلادرنگ پشتیبانی قوی از یادگیری انتقالی آموزش و پیاده‌سازی آسان	مدل عمومی تشخیص اشیاست، نه تخصصی برای متن. عملکرد ضعیف‌تر در کلمات خمیده یا طولانی
RRDETR* [۱۲]	مبتنی بر معماری ترانسفورمر DETR [۲۸] با ماژول تشخیص متن زاویه‌دار	•تشخیص عالی برای متن‌های چرخیده و مورب. بدون نیاز به روش حذف غیر بیشینه. آموزش انتها–به–انتها .	نیازمند محاسبات زیاد و GPU قوی. همگرایی کندتر نسبت به شبکه‌های CNN و حافظه مصرفی بالا.
RBDETR* [۱۳]	نسخه‌ی پیشرفته تر معماری ترانسفورمر با نام Deformale- DETR [۲۹] با ماژول تشخیص متن با شکل دلخواه.	عملکرد عالی در متن‌های منحنی، فارسی یا هنری، مقاوم در برابر تغییر شکل و زاویه متن، تشخیص یکپارچه متون با جهت‌های مختلف.	نیازمند محاسبات بالا نیاز به تنظیم دقیق پارامترها برای متون با حروف کوچک، همگرایی سریع تر
Flornence-۲ [۳۵]	مبتنی بر معماری ترانسفورمر زبان–بینایی آموزش‌دیده روی میلیاردها جفت تصویر–متن؛ برای وظایف بینایی و زبانی از جمله خواندن متن از تصاویر.	تعمیم‌پذیری بسیار بالا در تشخیص متن، زیرنویس‌گذاری آموزش روی داده‌های عظیم چندزبانه عملکرد قوی در OCR بدون آموزش مجدد	مدل بسیار سنگین و پرهزینه از نظر سخت‌افزاری. تنظیم و یادگیری مجدد دشوار. محدودیت در دسترسی به تنظیم کد تشخیص متن.
Qwen۲/۵-VL [۳۶]	مدل چند وجهی زبان–بینایی با پشتیبانی از خواندن متن از تصاویر و درک تصویر	قابلیت قوی در موقعیت یابی و فهم متن در تصویر توانایی تفسیر متون در صحنه‌های پیچیده پشتیبانی چندزبانه متن‌باز و قابل توسعه.	حجم زیاد مدل و تأخیر در پردازش. نیازمند ترکیب با آشکارساز جداگانه برای متون متراکم.

٤- نتایج تجربی

در این بخش، ارزیابی جامعی از عملکرد چندین معماری پیشرفته موقعیت‌یابی متن، شامل CRAFT [۱۱]، RRDETR [۱۲]، RBDETR [۱۳]، YOLOv۸ [۲۴]، Florence-۲ [۳۵] و Qwen۲/۵VL [۳۶] روی مجموعه داده‌ی FATD ارائه می‌شود. توزیع تصاویر آزمایشی مورد استفاده برای ارزیابی هر یک از این مدل‌ها مطابق با جدول ۲ است.

٤-۱جزئیات اجرایی

کلیه مدل‌های مورد استفاده در این پژوهش بر روی سیستمی مجهز به پردازنده گرافیکی NVIDIA-RTX-۳۰۹۰ آموزش داده و ارزیابی شده‌اند. جهت تضمین مقایسه‌ای عادلانه، مدل‌های انتخابی مورد استفاده در این مقاله، همگی در شرایط یکسان و با استفاده ازمجموعه داده‌های مشابه آموزش دیده‌اند. ارزیابی مدل‌ها در این مقاله فقط با استفاده از مجموعه داده ارزیابی ارائه شده در جدول ۲ انجام گرفته است و مدل‌ها روی داده‌ای غیر از فارسی آموزش دیده‌اند. عملکرد مدل‌ها با استفاده از معیار ارزیابی ICDAR۱۵، با حد آستانه پارامتر اشتراک به اجتماع[1] (IoU) بیشتر یا مساوی ۵۰ درصد ارزیابی می‌شوند [۴۳]. برای اطلاعات بیشتر درمورد جزییات اجرایی و پارامترهای مدل های مورد استفاده در این ارزیابی به جدول ۳ رجوع نمایید.

٤-۲ معیارهای ارزیابی موقعیت یابی متن

ارزیابی عملکرد مدل‌های از پیش آموزش دیده برای موقعیت یابی اشیاء متنی،معمولاً بر اساس پروتکل‌های ارزیابی استاندارد ارائه شده در [۴۳] و با معیارهای دقت، فراخوانی و میانگین هارمونیک صورت می‌گیرد. برای

جدول ۲: مشخصات آماری مجموعه داده (FATD)، طراحی‌شده برای موقعیت‌یابی متن فارسی. مقایسه مدل‌های بخش نتایج این مقاله روی داده ارزیابی معرفی‌شده.

انواع متن فارسی مجموعه داده		تعداد تصاویر	
		داده آموزش	داده ارزیابی
متن از محیط داخلی	محصولات فروشگاهی	۵۹	۱۹
	راهروها	۲۷۵	۹۱
متن از محیط خارجی	تابلوی مغازه‌ها	۹۲۴	۳۰۸
	تابلوهای خیابان	۳۹	۱۳
	دیوار نوشته‌ها	۵۱	۱۷
	تابلوهای راهنمایی رانندگی	۱۵۶	۵۲
کل		۱۵۰۴	۵۰۰

محاسبه دقت و فراخوانی، معیار اصلی ارزیابی، نسبت اشتراک به اجتماع (IoU) در نظر گرفته می‌شود. مقدار IoU برای هر کادر محصورکننده پیش‌بینی شده توسط مدل (D) و کادر محصورکننده داده‌ی مبنا (G) به شرح زیر محاسبه می‌شود:

$$IoU = \frac{Area(G \cap D)}{Area(G \cup D)} \qquad (۱)$$

در این رابطه، G نشان‌دهنده کادر محصور کننده داده مبن او D نشان‌دهنده کادر محصورکننده داده موقعیت‌یابی‌شده با استفاده از مدل یادگیری عمیق است.

معیار IoU درحوزه موقعیت‌یابی اشیاء مبتنی به طور گسترده برای ارزیابی دقت موقعیت‌یابی‌ها به کار می‌رود . برای آنکه یک موقعیت‌یابی به عنوان موقعیت‌یابی صحیح در نظر گرفته شود، باید دارای مقدار IoU

1. Intersection Over Union

جدول ۲: جزییات اجرایی و پارامترهای مدل‌های مورد استفاده در ارزیابی مدل‌هایی که با علامت * مشخص شده‌اند به صورت محلی آموزش داده شده‌اند.

مدل	منبع کد	وزن‌های اولیه (داده پیش آموزش)	یادگیری انتقالی	شرط پایان	رسم‌الخط مجموعه داده
[۱۱] CRAFT	لینک ۱	SynthText	ICDAR۱۳+MLT۱۷ICDAR	عدم بهبود در *H-mean*	عمدتاً انگلیسی و عربی
[۲۴] YOLOv۸*	لینک ۲	[۶۸] COCO	ICDAR۱۷	عدم بهبود در *H-mean*	عمدتا انگلیسی و عربی
[۱۲] RRDETR*	لینک ۲	SynthText	ICDAR۱۷MLT	عدم بهبود در *H-mean*	عمدتا انگلیسی و عربی
[۱۳] RBDETR*	لینک ۲	SynthText	ICDAR۱۷MLT	عدم بهبود در *H-mean*	عمدتا انگلیسی و عربی
[۳۵] Flornence-۲	لینک ۳	مجموعه دادگان Florence	داده‌های چندزبانه تصویری در مقیاس بزرگ	عدم بهبود خطای اعتبارسنجی	چندزبانه
[۳۶] Qwen۲/۵-VL	لینک ۴	مدل زبانی Qwen۲	داده‌های چندزبانه تصویری حجیم	عدم بهبود خطای اعتبارسنجی	چندزبانه

1. https://github.com/clovaai/CRAFT-pytorch
2. https://github.com/zobeirraisi/FATD
3. https://huggingface.co/microsoft/Florence-2-base
4. https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

بزرگتر یا مساوی ۰/۵ باشد. دقت (P) و فراخوانی (R) بر اساس روابط زیر محاسبه می‌شوند

$$P = \frac{TP}{TP + FP} \quad (۲)$$

$$R = \frac{TP}{TP + FN} \quad (۳)$$

در این روابط، TP[1] تعداد موقعیت‌یابی‌های صحیح مثبت، FP[2] تعداد موقعیت‌یابی‌های نادرست مثبت و FN[3] تعداد موقعیت‌یابی‌های نادرست منفی را نشان می‌دهند. معیار $H\text{-}mean$ نیز به عنوان میانگین هارمونیک دقت و فراخوانی به صورت زیر محاسبه می‌شود

$$H\text{-}mean = ۲ \times \frac{P \times R}{P + R} \quad (۴)$$

٤-۳ نتایج کمی

در این پژوهش، ابتدا عملکرد مدل‌های آموزشدیده منتخب، با استفاده از مجموعه آزمون منتخب از FATD ارزیابی و مقایسه شده است. برای اطمینان از منصفانه بودن مقایسه، کلیه مدل‌های CNN و ترانسفورمری (به جز مدل‌های زبان-بینایی) در ابتدا با استفاده از مجموعه داده‌ی مصنوعی [۵۰] آموزش دیده، سپس با استفاده از مجموعه داده‌ی ICDAR۱۷ [۵۲] تنظیم دقیق شده‌اند. اما مدل‌های VLM با آموزش پیش‌فرض و بدون تنظیمات اضافی استفاده شده‌اند. لازم به ذکر است، هیچکدام از این مدل‌ها با تصاویر فارسی مجموعه داده‌های معرفی شده‌ی این مقاله آموزش ندیده و تنظیم دقیق نشده‌اند. نتایج کمی این مقایسه‌ها در جدول ۴ درج شده است.

نتایج حاصل نشان می‌دهد که با وجود استفاده ازمجموعه داده‌های چندزبانه غیر فارسی ولی حاوی متن عربی (که دارای کاراکترهای مشابه با زبان فارسی است) برای آموزش، مدل‌های منتخب بر روی مجموعه داده‌ی معرفی‌شده عملکرد قابل قبولی ازخودنشان می‌دهند. این امر نشان‌دهنده قابلیت تعمیم‌پذیری مدل‌های موقعیت‌یابی متن روی مجموعه داده‌های ناشناخته با رسم‌الخط‌های متفاوت اما دارای کاراکترهای تقریبا مشابه است. در میانم دل‌های مورد بررسی، مدل مبتنی بر ترانسفورمر [۱۳] با معیار میانگین هارمونیک برابر با ۶۴/۶۵٪ بهترین عملکرد را روی مجموعه داده‌ی FATD داشت. مدل‌های زبان-بینایی با وجود اینکه روی یک مجموعه داده‌ی جامع از تصاویر مختلف آموزش دیده و با تصاویر متنی تنظیم دقیق نشده بودند، عملکرد مناسبی از خود نشان دادند، که انتظار می‌رود با انتقال یادگیری و تنظیم دقیق این مدل روی مجموعه داده‌ی متنی مشابه می‌توان به دقت بالاتری دست یافت. با وجود اینکه مدلهای ترانسفورمری بهترین عملکرد را از لحاظ دقت داشته‌اند، اما از سرعت پردازش مناسبی برخوردار نبوده و حتی بر روی یک پردازنده گرافیکی قوی نیز کمترین فریم بر ثانیه[4] (FPS) را داشته‌اند. بهترین عملکرد از لحاظ سرعت را مدل‌های CNNمحور به خود اختصاص داده و در نتیجه گزینه‌های مناسبی برای کاربردهایی هستند که سرعت در آنها مهم است. همانطور که انتظار می‌رفت مدل‌های VLMمحور به دلیل پیچیدگی معماری،نیازمند توان محاسباتی بالا بوده و از سرعت کمی برخوردارند و برای کاربردهای بلادرنگ قابل استفاده نیستند.

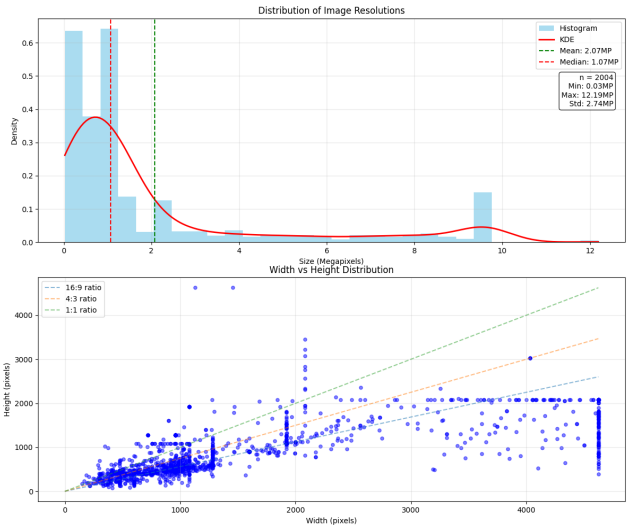
٤-٤ نتایج کیفی

برای ارزیابی عملکرد مدل‌ها در شرایط چالش‌برانگیز، آزمایش‌هایی بر روی مدل منتخب و با استفاده از نمونه‌های مختلف تصاویر از مجموعه FATD انجام گرفت. نتایج این آزمایشات کیفی با نتایج کمی ارائه‌شده در جدول ۴ همخوانی داشته و نشانم ی‌دهد که مدل ترنسفورمری RBDETR قادر به شناسایی دقیق‌تر نمونه‌های متن فارسی است، همان‌طورکه در شکل ۷ نشانداده شده است.

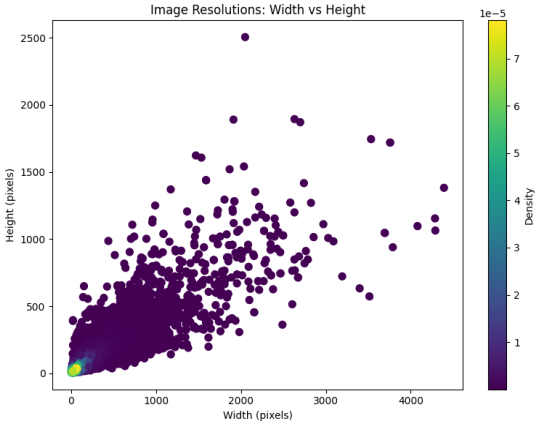
با اینکه مدل‌های مورد استفاده از بهترین و شناخته ترین مدل‌های یادگیری عمیق بوده و نتایج خوبی روی مجموعه داده معرفی‌شده داشته‌اند، اما این مدل‌ها دارای خطاهایی نیز هستند که در شکل ۷ با رنگ زرد برجسته شده‌اند. همان‌طور که در این شکل می‌توان دید ، تمامی این مدل‌ها تقریباً روی متون با انسداد حتی جزیی (شکل ۷- ردیف آخر) نیز موفق به موقعیت‌یابی نشده‌اند-، مثلاً کلماتی مانند "استارت، بزنید، کاغذ، کفپوش، سنگ، مصنوعی" به‌دلیل انسداد موقعیت‌یابی نشده‌اند و این حقیقت لزوم توجه محققین به این چالش را در آینده نشان می‌دهد.

1. True Positive
2. False Positive
3. False Negative

4. Frame per Second



شکل ۵: توزیع آماری وضوح و ابعاد تصاویر مورد استفاده در مجموعه داده FATD. نمودار بالا، تابع چگالی احتمال وضوح، و نمودار پایین تابع چگالی احتمال نسبت ارتفاع به عرض تصاویر موجود در مجموعه داده.



شکل ۶: توزیع آماری نسبت ارتفاع به عرض نواحی متنی (کلمات) در مجموعه داده FATD.

مدل‌های بزرگ زبانی فقط قادر به موقعیت‌یابی خط متنی کامل در تصویر شده‌اند و در تشخیص کلمات به صورت جداگانه دچار مشکل بوده‌اند (شکل ۷–الف) علاوه بر این، این مدل‌ها کادر مستطیلی را در خروجی هنگام موقعیت‌یابی نمایش می‌دهند که برای متون نوشته‌شده افقی مناسب بوده، ولی برای متن با چرخش یا مورب کارایی ندارند. (شکل ۷–الف).

این مشکل برای مدل YOLO نیز مشهود است (شکل ۷–ج). مدل YOLO چون دارای وزن پیش‌آموزش دیده شده روی مجموعه داده‌های عمومی نیز هست، با وجود انتقال یادگیری با متون عربی و انگلیسی، در تصویر ردیف بالا نماد هواپیما و پیکان را نیز موقعیت‌یابی کرده است که یک خطا می‌باشد. علاوه براین، این مدل در موقعیت‌یابی بسیاری از کاراکترها ("دا"، "دک") و هم چنین کلمات ("با اطمینان") دچار خطا شده است.

مدل RBDETR نسبت به دیگر مدل‌هاعملکرد بهتری داشته است، اما این مدل همانطور که در شکل ۷– ب مشاهده می‌شودنیز در موقعیت‌یابی کلمه "با اطمینان" و هم چنین در کلمه "تاپ" بیش از کلمه را مکان‌یابی کرده است.

از آنجایی که این مدل‌ها روی متون فارسی آموزش ندیده اند، برخی از این خطاها را می‌توان با انتقال یادگیری و آموزش روی مجموعه داده

(الف) (ب) (ج) (د) (ه)

شکل ۷: مقایسه کیفی عملکرد خروجی مدل‌های مورد ارزیابی با تصاویر FATD که در آن (الف) Qwen۲/۵-VL، (ب) RBDETR، (ج) YOLOv۸، (د) CRAFT و (ه) تصویر برچسب را نشان می‌دهند. قاب های مستطیلی زرد رنگ دور متن‌ها، خطای خروجی هر مدل را نمایش می‌دهد.

جدول ۳: نتایج کمی برای مدل های مورد ارزیابی روی مجموعه داده FATD.

مدل	مجموعه داده FATD			سرعت
	دقت	فراخوانی	میانگین هارمونیک	(فریم بر ثانیه)
[۱۱] CRAFT	۶۵٫۱۶	۶۰٫۰۸	۶۲٫۵۲	۲۸
[۲۴] YOLOv۸	۶۲٫۱۱	۵۸٫۳۴	۶۰٫۱۶	۴۳
[۱۲] RRDETR	۶۵٫۴۲	۶۲٫۲۱	۶۳٫۷۷	۸٫۸
[۱۳] RBDETR	۶۶٫۹۲	۶۴٫۴۱	۶۴٫۶۵	۸٫۵
[۳۵] Florence-۲	۵۸٫۱۲	۵۴٫۲۳	۵۶٫۱۱	۰٫۳
[۳۶] Qwen۲/۵-VL	۶۰٫۷۲	۵۸٫۹۸	۵۹٫۸۴	۰٫۲

فارسی مثل FATD معرفی‌شده در این مقاله تا حدودی برطرف کرد. اما برای رفع خطاهایی مثل انسداد متن نیاز به طراحی مدل کاملاً جدید و با معماری مخصوص می‌باشد.

باوجود نتایج امیدوارکننده، همانطور که از شکل ۷ پیداست، چالش‌هایی همچون وضوح پایین متون، جهت‌گیری نامناسب و انسداد حتی جزیی در متن درشرایط واقعی، مانع از دستیابی به عملکرد مطلوب شده است. لذا، لازم است تحقیقات آتی بر رفع این موانع متمرکز شوند.

٤–٥ بحث و کارهای آتی

موقعیت‌یابی و شناسایی متن با چالش‌ها یقابل توجهی از جمله فرآیند زمان‌بر و پرهزینه نشانه‌گذاری مواجهاست. پیشرفت‌های اخیر در حوزه هوش مصنوعی، مانند مدل‌های تولید تصویر و مدل‌های زبانی، امکان تولید تصاویر بر اساس دستورات متنی را فراهم و راهکاری بالقوه برای کاهش نیاز به داده‌های نشانه‌گذاری‌شده دستیارائه می‌دهد [۶۹] تا [۷۲].

یکی دیگر از روش‌های مقابله با چالش نشانه‌گذاری، خودکارسازی این فرآیند با استفاده از تکنیک‌های یادگیری خودنظارتی و مدل‌های پیشرفته مانند مدل ارائه‌شده در [۷۳] است. با این حال، موقعیت‌یابی و شناسایی متن فارسی و همچنین متن در تصاویر واقعی همچنان با چالش‌هایی مانند تغییرات جهت‌گیری متن، انسداد توسط اشیاء دیگر و افت کیفیت تصویر مواجه است. این چالش‌ها همچنان به عنوان مسائل حل نشده در حوزه بینایی ماشین مطرح هستند. برای غلبه بر اینچالش‌ها، می‌توان از تکنیک‌هایی مانند تقویت داده‌ها، ترکیب‌پذیری [۷۴]، ترانسفورمرهای خودرمزگذار ماسک‌شده [۷۵] به همراه ماژول‌های هوش مصنوعی پیشرفته استفاده کرد.

در این مقاله مدل های مورد استفاده، اغلب روی مجموعه دادگان انگلیسی و چندزبانه که شامل فقط متون عربی می‌باشند، آموزش دیده و هیچگونه آموزش یا پیش-آموزشی روی رسم الخط یا زبان فارسی انجام نشده است؛ با این حال، این مدل‌ها تقریباً عملکرد قابل قبولی را در موقعیت‌یابی خط فارسی ارائه داده‌اند؛ لذا می‌توان انتظار داشت که با آموزش یا انتقال یادگیری روی مجموعه داده فارسی مانند مجموعه داده معرفی شده در این مقاله بتوان دقت و عملکرد این مدل‌ها را افزایش دادو این مقوله در برنامه تحقیقاتی نگارندگان این مقاله است.

موقعیت یابی دقیق کلمات متون فارسی مخصوصاً وقتی با خطوط نستعلیق شکسته و ثلث نوشته می‌شوند و ممکن است دچار خمیدگی و در هم تنیدگی‌های پیچیده‌ای باشند که به حاشیه‌نویسی چند ضلعی با رئوس زیاد نیاز دارند، در این مقاله به آنها پرداخته نشده است. از آنجایی که این کلمات به وفور در نوشته‌های فارسی در تصاویر ظاهر می‌شوند، لازم است محققان در آینده بیشتر به این موضوع توجه داشته باشند.

موقعیت‌یابی و شناسایی همزمان متن در تصاویر (E۲ESTDR) که شامل تشخیص موقعیت مکانی و خواندن متن است، یکی از حوزه‌های فعال در بینایی کامپیوتر است که در زبان‌های دیگر مورد مطالعه قرار گرفته است. این مقوله برای متن فارسی در تصاویر واقعی، با توسعه و بهبود روش‌های موجود، از برنامه‌های آینده ما خواهد بود.

۵- نتیجه‌گیری

این مقاله با هدف رفع چالش‌های مکان‌یابی متن فارسی در تصاویر طبیعی، ابتدا مجموعه‌داده جامع FATD (با ۲۰۰۰ تصویر چالش‌برانگیز حاوی تنوع بالای فونت، اندازه، زاویه، نور و پیچیدگی صحنه) را به‌عنوان منبعی استاندارد معرفیو شش مدل پیشرفته یادگیری عمیق را با استفاده از این منبع داده ارزیابی نموده است. ارزیابی سیستماتیک این شش مدلکه در سه دسته‌یCNNها (CRAFT و YOLOv۸)، ترانسفورمرها (RRDETR و RBDETR) و مدل‌های زبان-بینایی (Florence-۲ و Qwen۲/۵VL) طبقه‌بندی شده‌اند، نشان می‌دهد که ترنسفورمرها با دقت حدود۶۵درصد (براساس معیار $H\text{-}mean$) بهترین عملکرد را به قیمت هزینه محاسباتی بالا دارا هستند. CNNها با حفظ دقتی رقابتی و سرعت پردازش مطلوب، گزینه‌ی بهینه برای کاربردهایبا تاخیر کمتر محسوب می‌شوند. عملکرد نسبتاً خوب مدل‌های زبان-بینایی علیرغم آموزش محدود، پتانسیل بالای آن‌ها را در این حوزه نشان می‌دهد. این مطالعه با ارائه‌ی منبع داده‌ای معتبر و ارزیابی مقایسه‌ای، بستری اساسیبرای توسعه‌ی راهکارهای کارآمدتر مکان‌یابی متن فارسی فراهم ساخته است.بااینحال، هنوز دقت به دست آمده بسیار دور از انتظار و پایین تر از دقت این الگوریتم ها در مجموعه داده‌هایارزیابی استاندارد است، در نتیجه برای دستیابی به دقت بالاتر در موقعیت‌یابی و شناسایی متن فارسی در تصاویر واقعی، نیاز به توسعه و استفاده از معماری یادگیری عمیق اختصاصی با تمرکز بر پیچیدگی‌های متن فارسی احساس می‌شود.

مراجع

[1] Y. Zhu, C. Yao, and X. Bai, "Scene text detection and recognition: Recent advances and future trends," *Front. Comp Sci.*, vol. 10, no. 1, pp. 19-36, 2016.

[2] H. Lin, P. Yang, and F. Zhang, "Review of scene text detection and recognition," *Arch. Comput. Methods Eng*, pp. 1-22, 2019.

[3] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, *Text Detection and Recognition in the Wild: A Review*, arXiv: arXiv:2006.04305, 2020.

[4] Z. Raisi and J. Zelek, "Text detection and recognition for robot localization," *J. Electr. Comput. Eng. Innov. JECEI*, vol. 12, no. 1, pp. 163174, 2024.

[5] X. Han, J. Gao, C. Yang, Y. Yuan, and Q. Wang, "Focus entirety and perceive environment for arbitrary-shaped text detection," *IEEE Trans. Multimed.*, 2024.

[6] J. Xu *et al.*, "FSANet: Feature shuffle and adaptive channel attention network for arbitrary shape scene text detection," *Neurocomputing*, Article ID: 129443, 2025.

[7] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. on Comp. Vision*, pp. 1457-1464, Barcelona, Spain, 6-13 Nov. 2011.

[8] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven, " PhotoOCR: Reading text in uncontrolled conditions," in *Proc. IEEE Int. Conf. on Comp. Vision*, pp. 785-792, Sydney, Australia, 1-8 Dec. 2013.

[9] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, pp. 5551-5560, Honolulu, HI, USA, 21-26 Jul.2017.

[10] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," *IEEE Trans Image Process*, vol. 27, no. 8, pp. 3676-3690, Apr. 2018.

[11] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, pp. 9365-9374, Long Beach, CA, USA, 16-17 Jun. 2019.

[12] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, and J. S. Zelek, "Transformer-Based text detection in the wild," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pp. 3162-3171, Nashville, TN, USA,19-25 Jun. 2021.

[13] Z. Raisi, G. Younes, and J. Zelek, "Arbitrary shape text detection using transformers," in *Proc. 26th Int. Conf. on Pattern Recognition*, pp. 3238-3245, Montreal, Canada,2022.

[14] Z. Raisi and J. Zelek, "Visual place recognition from end-to-end semantic scene text features," *Front. Robot. AI*, vol. 11, Article ID: 1424883, Sept. 2024.

[۱۵] ز. حیدران داروقه امنیه، س. م. رستگار فاطمی، م. رستگارپور و گ. آقایی قزوینی، "افزایش دقت شبکه‌های عصبی کانولوشنی مبتنی بر مدل چهار-جریان با فیلترهای پردازش تصویر و نگاشت خطی‌ساز فضای عدم تشابه"، *نشریه روش‌های هوشمند در صنعت برق*، سال ۱۶، شماره ۶۱، صص. ۲۸-۱، بهار ۱۴۰۴.

[۱۶] م. روحی، ج. مظلوم، م. ع. پورمینا و ب. قلمکاری، "طبقه‌بندی سکته مغزی بر اساس روش یادگیری عمیق در سیستم تصویربرداری ریزموجی از مغز"، *نشریه روش‌های هوشمند در صنعت برق*، سال ۱۵، شماره ۵۷، صص. ۱۳۲-۱۲۱، بهار ۱۴۰۳.

[۱۷] ف. علی‌مرادی، ف. رحمانی، ل. ربیعی، م. خوانساریو م. مازوچی، "ساخت مجموعه داده تصاویر برای تشخیص و بازشناسی متن در تصاویر، "*فصلنامه اطلاعات و ارتباطات ایران*، سال ۱۴، شماره ۵۳، صص. ۹۵-۷۸، پاییز-زمستان ۱۴۰۱.

[18] S. Kheirinejad, N. Riaihi, and R. Azmi, "Persian text-based traffic sign detection with convolutional neural network: A new dataset," in *Proc. 10th Int. Conf. on Computer and Knowledge Engineering*, pp. 060-064, Mashhad, Iran, 29-30 Oct. 2020.

[19] A. Fateh, M. Rezvani, A. Tajary, and M. Fateh, "Persian printed text line detection based on font size," *Multimed. Tools Appl.*, vol. 82, no. 2, pp. 2393-2418, Jan. 2023.

[20] M. Rahmati, M. Fateh, M. Rezvani, A. Tajary, and V. Abolghasemi, "Printed Persian OCR system using deep learning," *IET Image Process.*, vol. 14, no. 15, pp. 3920-3931, Dec. 2020.

[21] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, pp. 2963-2970, San Francisco, CA, USA, 13-18 Jun. 2010.

[22] H. Chen, *et al.*, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Proc. IEEE Int. Conf. on Image Processing*, pp. 2609-2612, Barcelona, Spain,6-13 Nov. 2011.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. in Neural Info. Process. Sys.*, pp. 91-99, Montreal, Canada, 7-12 Dec.2015.

[24] M. Yaseen, What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector, arXiv preprint arXiv:2408.15857, 2024.

[25] W. Liu, *et al.*, "SSD: Single shot multibox detector," in *Eur. Conf. on Comp. Vision*, Springer, pp. 21-37, 2016.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, 2015, pp. 3431–3440.
[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. on Comp. Vision*, pp. 2961-2969, 2017.
[28] N. Carion, *et al*., *End-to-End Object Detection with Transformers*, arXiv Preprint arXiv200512872, 2020.
[29] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, *Deformable DETR: Deformable Transformers for End-to-End Object Detection*, arXiv Preprint arXiv201004159, 2020.
[30] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Proc. Eur. Conf. on Comp. Vision*, Springer, pp. 56-72, 2016.
[31] S. Long, *et al.*, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in *Proc. Eur. Conference. on Computer Vision*, pp. 20-36, 2018.
[32] D. Deng, H. Liu, X. Li, and D. Cai, "Pixellink: Detecting scene text via instance segmentation," in *Proc. AAAI Conf. on Artif. Intell.*, 2018.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 234-241, 2015.
[34] L. Yuan, *et al.*, *Florence: A New Foundation Model for Computer Vision*, arXiv preprint arXiv:2111.11432, 2021.
[35] B. Xiao, *et al.*, *Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks*, arXiv preprint arXiv:2311.06242, 2023.
[36] S. Bai, *et al.*, *Qwen2.5-VL Technical Report*, arXiv preprint arXiv:2502.13923, 2025
[37] J. Bai, *et al.*, *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond*, arXiv preprint arXiv:2308.12966, 2023.
[38] A. Hurst, *et al., GPT-4o System Card*, arXiv preprint arXiv:2410.21276, 2024.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *in Proc. Int. Conf. on Learning Representations*, 2015.
[40] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in*Proc. Medical Image Computing and Computer-Assisted Intervention*, pp. 234-241, 2015.
[41] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, arXiv PreprintarXiv1506.02640, 2016.
[42] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *Proc. Int. Conf. on Document Anal. and Recognition*, pp. 1484-1493, 2013.
[43] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. Int. Conf. on Document Anal. and Recognition*, pp. 1156-1160, 2015.
[44] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, *Coco-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images*, arXiv preprint. arXiv160107140, 2016.
[45] A. Singh, et al., "TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 8802-8812, 2021.
[46] C. K. Ch'ng and C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in *Proc. IAPR Int. Conf. on Document Anal. and Recognition*, pp. 935-942, Kyoto, Japan, 9-11 Nov. 2017.
[47] L. Yuliang, J. Lianwen, Z. Shuaitao, and Z. Sheng, *Detecting Curve Text in the Wild: New Dataset and New Solution*, arXiv preprint arXiv:1712.02170, 2017.
[48] S. Long, S. Qin, Y. Fujii, A. Bissacco, and M. Raptis, "Hierarchical text spotter for joint text spotting and layout analysis," in *Pro. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, pp. 903-913, 2024.
[49] T. Kil, S. Kim, S. Seo, Y. Kim, and D. Kim, "Towards unified scene text spotting based on sequence generation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 15223–15232.
[50] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic Data for Text Localisation in Natural Images," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition*, 2016.
[51] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, *Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition*, arXiv PreprINT. ArXiv14062227, 2014.
[52] M. Iwamura, *et al.*, "ICDAR2017 robust reading challenge on omnidirectional video," in *Proc. IAPR Int. Conf. on Document Anal. and Recognition*, 2017, pp. 1448–1453.
[53] C. -K. Chng, *et al*., ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text (RRC-ArT), arXiv preprint arXiv1909.07145, 2019.
[54] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognit.*, 2012, pp. 1083–1090.
[55] W. Wu, *et al.*, "ICDAR 2023 competition on video text reading for dense and small text," in *Proc. Int. Conf. on Document Analysis and Recognition*, pp. 405-419, 2023.
[56] Z. Wan, J. Zhang, L. Zhang, J. Luo, and C. Yao, "On vocabulary reliance in scene text recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 11425-11434, 2020.
[57] R. Zhang *et al.*, "ICDAR 2019 robust reading challenge on reading Chinese text on signboard," in *Proc. International Conf.e on Document Analysis and Recognition* 2019, pp. 1577–1581.
[58] M. Tounsi, I. Moalla, A. M. Alimi, and F. Lebouregois, "Arabic characters' recognition in natural scenes using sparse coding for feature representations," in *Proc. 13th Int. Conference on Document Analysis and Recognition*, pp. 1036-1040, 2015.
[59] M. Tounsi, I. Moalla, and A. M. Alimi, "ARASTI: A database for Arabic scene text recognition," in *Proc. 1st Int. Workshop on Arabic Script Analysis and Recognition*, pp. 140-144, 2017.
[60] M. Jain, M. Mathew, and C. Jawahar, "Unconstrained OCR for Urdu using deep CNN-RNN hybrid networks," in *Proc. 4th IAPR Asian Conf. on Pattern Recognition,* pp. 747-752, 2017.
[61] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Advances in Neural Information Processing Systems*, pp. 3856-3866, 2017.
[62] A. Rahman, A. Ghosh, and C. Arora, "UTRNet: High-Resolution Urdu Text Recognition in Printed Documents," in *Proc. Int. Conf. on Document Analysis and Recognition*, pp. 305-324, 2023.
[63] E. Shabaninia, F. Eslami, A. Afkari-Fahandari, and H. Nezamabadi-pour, "SUT: a new multi-purpose synthetic dataset for Farsi document image analysis," in *Proc. 13th Int. Conf. on Computer and Knowledge Engineering*, pp. 253-258, 2023.
[64] F. Asadi-Zeydabadi, E. Shabaninia, H. Nezamabadi-Pour, and M. Shojaee, "Farsi optical character recognition using a transformer-based model," in *Proc. 13th Int. Conf. on Computer and Knowledge Engineering*, pp. 293-299, 2023.
[65] M. Mosannafat, F. Taherinezhad, H. Khotanlou, and E. Alighardash, "Farsi text detection and localization in videos and images," in *Proc. 9th Iranian Joint Congress on Fuzzy and Intelligent Systems*, 6 pp., Bam, Iran, 2-4 Mar. 2022.
[66] A. Salmasi and E. Kabir,"Farsi text in scene: A new dataset," in *Proc. 13th Int. Conf. on Computer and Knowledge Engineering*, pp. 510-514, Mashhad, Iran, Nov. 2023.
[67] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. of the 27th ACM Int. Conf. on Multimedia*, New York, NY, USA. doi: 10.1145/3343031.3350535.
[68] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context" in *Proc. Euro. Conf. on Comp.Vision*, pp. 740-755*,* Milan, Itay, 29 Sept.-4 Oct. 2014.
[69] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical Text-Conditional Image Generation with Clip Latents*, arXiv Preprint, arXiv220406125, 2022.
[70] J. Achiam *et al.*, *GPT-4 Technical Report*, arXiv preprint arXiv230308774, 2023.
[71] D. Guo,et al., DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv Preprint arXiv250112948, 2025.
[72] G. Team, et al., Gemini: A Family of Highly Capable Multimodal Models, arXiv Preprint, arXiv231211805, 2023.
[73] A. Kirillov *et al.*, "Segment anything," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 4015-4026, 2023.
[74] A. Kortylewski, Q. Liu, H. Wang, Z. Zhang, and A. Yuille, "Combining compositional models and deep networks for robust object classification under occlusion," in *Proc. IEEE Winter Conf. on Applications of Computer Vision*, pp. 1333-1341, 2020.
[75] Z. Raisi and J. Zelek, "Occluded text detection and recognition in the wild," in *Proc. 19th Conf. on Robots and Vision*, pp. 140-150, Toronto, Canada, May 2022.

زبیر رئیسی تحصیلات خود را در مقاطع کارشناسی مهندسی برق- الکترونیک و کارشناسی ارشد مهندسی برق- مخابرات به‌ترتیب در سال‌های ۱۳۸۸ و ۱۳۹۰ از دانشگاه سیستان و بلوچستان و در مقطع دکتری مهندسی طراحی سیستم در سال ۱۴۰۱ از دانشگاه واترلوی کانادا به پایان رسانده است. پس از اتمام دوره دکتری، نام‌برده به عنوان پژوهشگر پسادکتری در دانشگاه واترلو فعالیت داشته و در چندین پروژه علمی- صنعتی با

شرکت‌های برجسته‌ای از جمله Apple، Baltimore Orioles و ATS Automation همکاری داشته است. هم‌اکنون، ایشان استادیار دانشکده مهندسی دریا در دانشگاه دریانوردی و علوم دریایی چابهار می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: بینایی کامپیوتر، هوش مصنوعی، موقعیت یابی و تشخیص اشیا و متن از تصاویر طبیعی، تشخیص اشیا از تصاویر ماهواره ای، و کالیبراسیون دوربین می‌باشد.

رسول دامنی، استادیار دانشگاه دریانوردی و علوم دریایی چابهار، دارای مدرک دکترای مهندسی برق در گرایش مخابرات سیستم از دانشگاه صنعتی شریف (اخذ شده در سال ۱۳۹۳) است. حوزه تخصصی وی در زمینه سیستم‌های مخابراتی با تمرکز بر مخابرات نوری و پردازش سیگنال‌های نوری است. زمینه‌های مورد علاقه پژوهشی وی علاوه بر سیستم‌های مخابرات نوری، شامل سیستم های مخابراتی زیر آب و هوش مصنوعی (بینایی ماشین) است و در این حوزه‌ها دارای سوابق پژوهشی و انتشار مقالات علمی است.

اسماعیل سارانی، دکترای مهندسی برق- قدرت را در سال ۱۳۹۶ از دانشگاه تهران دریافت نمود. ایشان هم‌اکنون استادیار گروه مهندسی الکترونیک و مخابرات دریایی دانشگاه دریانوردی و علوم دریایی چابهار می‌باشد. زمینه‌های مورد علاقه ایشان شامل ماشین‌های الکتریکی، انرژی‌های نو، الکترونیک قدرت و کاربردهای هوش مصنوعی و پردازش تصویر در مهندسی برق است. وی تاکنون مقالات متعددی در این زمینه‌ها در مجلات علمی معتبر داخلی و بین‌المللی به چاپ رسانده و تحقیقاتش بر توسعه راهکارهای نوین در تقاطع مهندسی قدرت کلاسیک، فناوری‌های پیشرفته و کاربرد یادگیری عمیق در پردازش تصویر متمرکز می‌باشد.

ولی محمد نظرزهی حاد در سال ۱۳۹۵ موفق به اخذ درجه دکترا در رشته مهندسی برق (سیستم‌ها و کنترل) از دانشگاه نیو ساوت ولز استرالیا گردید. وی هم‌اکنون به عنوان استادیار در گروه مهندسی برق دانشگاه دریانوردی چابهار مشغول به فعالیت می‌باشد. زمینه‌های پژوهشی ایشان شامل کنترل غیرمتمرکز، سامانه‌های رباتیک خودمختار، سیستم‌های کنترل دریایی هوش مصنوعی و بینایی ماشین است. وی عضو انجمن مهندسان برق و الکترونیک (IEEE) بوده و تاکنون مقالات متعددی را در مجلات معتبر علمی و کنفرانس‌های بین‌المللی در حوزه‌های سیستم‌های کنترل، رباتیک و فناوری‌های حسگری هوشمند به چاپ رسانده است.