مقاله پژوهشی

تقویت فراتفکیک‌سازی تصاویر متنی توسط تضعیف عامدانه تابع زیان خوانش برای اعمال سخت‌گیری بیشتر بر شبکه فراتفکیک‌ساز

کمیل مهرگان، عباس ابراهیمی مقدم و مرتضی خادمی درح

چکیده: تصاویر متنی با وضوح پایین معمولاً باعث ایجاد خطاهای جدی در خوانش و بازیابی متن می‌شوند که این امر می‌تواند بر کارایی سیستم‌های خوانش متن، تأثیر منفی بگذارد. فراتفکیک‌سازی تصاویر متنی، به‌ویژه در شرایطی که تصاویر اولیه دارای تفکیک‌پذیری پایینی هستند، از عوامل کلیدی در بهبود دقت سیستم‌های خوانش متن است. روش‌های سنتی فراتفکیک‌سازی، هرچند در بهبود کیفیت تصاویر موفق بوده‌اند، اما همچنان در حفظ جزئیات دقیق حروف و ساختار متن با چالش مواجهند. در این پژوهش، روشی برای فراتفکیک‌سازی تصاویر متنی ارائه شده که با بهره‌گیری از بازخورد هوشمندانه توسط تضعیف عامدانه تابع زیان خوانش، سخت‌گیری بیشتری بر شبکه فراتفکیک‌ساز اعمال کرده تا به‌طور ویژه تصاویری تولید کند که در آن ساختار حروف به‌خوبی حفظ شده باشد. این تابع زیان، شبکه فراتفکیک‌سازی را وادار به بازسازی جزئیات ازدست‌رفته در تصاویر کرده و دقت سیستم‌های خوانش متن را به‌طور قابل توجهی بهبود می‌بخشد. نتایج تجربی نشان می‌دهند که این روش نه‌تنها به افزایش وضوح بصری تصاویر منجر می‌شود، بلکه کارایی و دقت سیستم‌های خوانش متن را حدود ۱۰ درصد نسبت به تصاویر اولیه بهبود می‌بخشد. این رویکرد جدید گامی مؤثر در جهت بهینه‌سازی فرایندهای خوانش متن از تصاویر با تفکیک‌پذیری پایین به شمار می‌رود.

کلیدواژه: **بازخورد هوشمندانه، تضعیف عامدانه تابع زیان، خوانش تصاویر متنی، فراتفکیک‌سازی.**

۱- مقدمه

خوانش دقیق حروف از روی تصاویر با استفاده از روش‌های خوانش متن (OCR)[1] یکی از نیازهای اصلی در بسیاری از کاربردها است. با این حال در مواجهه با تصاویر با وضوح پایین (LR)[2]، این فرایند با چالش‌های متعددی همراه است که اغلب به اشتباه در خوانش حروف منجر می‌شود [۱] و [۲]. در تصاویر با کیفیت پایین، بخش قابل توجهی از اطلاعات، به‌ویژه جزئیات و اطلاعات فرکانس بالا از بین می‌رود. این موضوع به‌طور مستقیم عملکرد سیستم‌های خوانش متن را تحت تأثیر قرار می‌دهد و

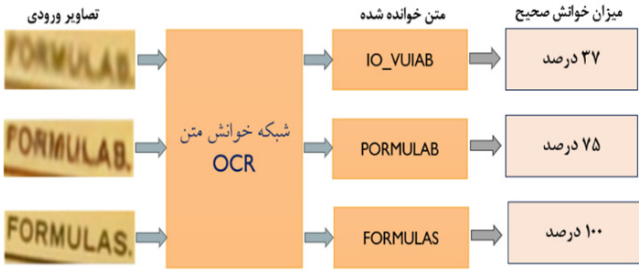
این مقاله در تاریخ ۳ دی ماه ۱۴۰۳ دریافت و در تاریخ ۱۷ فروردین ماه ۱۴۰۴ بازنگری شد.

کمیل مهرگان، دانشکده مهندسی برق، دانشگاه فردوسی مشهد، مشهد، ایران، (email: komail.mehrgan@mail.um.ac.ir).

عباس ابراهیمی مقدم (نویسنده مسئول)، دانشکده مهندسی برق، دانشگاه فردوسی مشهد، مشهد، ایران، (email: a.ebrahimi@um.ac.ir).

مرتضی خادمی درح، دانشکده مهندسی برق، دانشگاه فردوسی مشهد، مشهد، ایران، (email: khademi@um.ac.ir).

1. Optical Character Recognition
2. Low Resolution



شکل ۱: خوانش تصاویر با تفکیک‌پذیری‌های متفاوت.

دقت آنها را کاهش می‌دهد.

همان طور که در شکل ۱ نشان داده شده است، یک شبکه خوانش متن هنگام کار با تصاویر با وضوح‌های مختلف، دارای دقت متفاوتی است. استفاده از دوربین‌هایی با قابلیت ثبت تصاویر با وضوح بالا (HR)[3] به‌عنوان یک راهکار پیشنهاد شده است، اما این روش هزینه‌های بالایی را به همراه دارد. به همین دلیل، استفاده از روش‌های فراتفکیک‌سازی[4] تصاویر به‌عنوان یک راه‌حل کارآمد مطرح می‌شود [۳]. این تحقیق سعی دارد که با معرفی یک تابع زیان مبتنی بر تضعیف عامدانه خوانش برای اعمال سخت‌گیری بر شبکه فراتفکیک‌سازی به تقویت شبکه‌های فراتفکیک‌سازی بپردازد.

۲- پیشینه پژوهش

پیشینه تحقیق در زمینه فراتفکیک‌سازی تصاویر را می‌توان به سه دسته اصلی تقسیم کرد:

۱) روش‌های مبتنی بر بهینه‌سازی مسئله معکوس

۲) روش‌های مبتنی بر درون‌یابی

۳) روش‌های مبتنی بر یادگیری

از جمله روش‌های مؤثر در فراتفکیک‌سازی تصاویر متن به روش مبتنی بر بهینه‌سازی مسأله معکوس و روش‌های مبتنی بر درونیابی می‌توان به‌ترتیب به [۴] و [۵] اشاره کرد. روش‌های مبتنی بر یادگیری نیازمند امکانات سخت‌افزاری فراوان و داده‌های زیادی برای آموزش هستند. در سال‌های اخیر به دلیل فراهم‌آمدن این امکانات، توجه بیشتری به این روش‌ها جلب شده است؛ زیرا این روش‌ها توانسته‌اند پاسخ‌های دقیق‌تری را نسبت به دو روش دیگر ارائه دهند. برای نخستین بار دونگ[5] و همکاران در سال ۲۰۱۴ [۶] از شبکه‌های عصبی- کانولوشنی[6] (CNN) در

3. High Resolution
4. Super-Resolution
5. Dong
6. Convolutional Neural Network

جدول ۱: نمونه‌های مطرح از معماری‌های شبکه فراتفکیک‌سازی.

ردیف	معماری شبکه	یک نمونه مطرح	مراجع
۱	اتصالات خطی	FSRCNN	[۱۰] تا [۱۲]
۲	اتصالات باقیمانده	RFANet	[۱۳] تا [۱۵]
۳	اتصالات پیش‌رونده بازساز	LapSRN	[۱۶] و [۱۷]
۴	اتصالات بازگشتی	MemNet	[۱۸] و [۱۹]
۵	اتصالات مبتنی بر تمرکز	MADNet	[۲۰]
۶	اتصالات چندشاخه‌ای	DBFRN	[۲۱]
۷	اتصالات چگال متصل	SRDenseNet	[۲۲]
۸	اتصالات مدیریت تخریب چندگانه	SRNDNF	[۲۳]
۹	مدل شبکه‌های مولد مجادلانه	SRGAN	[۲۴] تا [۲۶]

طراحی شبکه‌های فراتفکیک‌ساز بهره بردند. شبکه پیشنهادی آنها تنها شامل سه لایه کانولوشنی بود. در ادامه، این معماری برای فراتفکیک‌سازی ویدئو گسترش یافت [۷]. در این روش، فریم‌های متوالی و نزدیک به هم یک ویدئو به عنوان ورودی شبکه در نظر گرفته شده و پس از پردازش در یک شبکه کانولوشنی دیگر، تصویر نهایی با وضوح بالا تولید می‌شود. هردایس و همکاران نیز با افزایش تعداد لایه‌های کانولوشنی، بهبود چشمگیری در کیفیت تصاویر بازسازی‌شده نشان دادند [۸].

معماری شبکه‌های فراتفکیک‌سازی مبتنی بر یادگیری بر اساس نحوه اتصال بلوک‌ها و لایه‌ها نقش مهمی در عملکرد این شبکه‌ها ایفا می‌کند و می‌توان آنها را به دسته‌های زیر تقسیم‌بندی کرد:

۱) اتصالات خطی[1]: ساده‌ترین نوع معماری، اتصالات خطی هستند که در آنها تنها یک مسیر جریان سیگنال وجود دارد و لایه‌های کانولوشنی به صورت متوالی پشت سر هم قرار گرفته‌اند. یکی از اولین نمونه‌ها در این زمینه، معماری شبکه SRCNN[2] [۹] بود. این شبکه از یک لایه نمونه‌افزایی[3] در ابتدای مسیر و چندین لایه کانولوشنی تشکیل شده که به صورت خطی به هم متصل هستند. نسخه سریع‌تر این مدل یعنی FSRCNN[4] از نمونه‌افزایی دیکانولوشنی در انتهای مسیر استفاده کرده که موجب افزایش سرعت و دقت عملکرد شبکه می‌شود.

۲) اتصالات باقیمانده[5]: در برخی معماری‌ها از اتصالات باقیمانده استفاده می‌شود که به شبکه اجازه می‌دهد اطلاعات را از لایه‌های ابتدایی به لایه‌های انتهایی منتقل کند؛ بدون اینکه این اطلاعات در طول مسیر از دست بروند. این نوع اتصالات به مقابله با مشکل محوشدگی گرادیان کمک می‌کنند و به شبکه اجازه می‌دهند تا عمق بیشتری داشته باشد. مثلاً در شبکه VDSR[6] با استفاده از تکنیک این اتصالات و یادگیری باقیمانده[7]، توانسته است سرعت همگرایی را بهبود بخشد.

۳) اتصالات پیش‌رونده بازساز[8]: در این نوع از اتصالات از تکنیکی استفاده می‌کنند که افزایش وضوح تصویر را به صورت مرحله‌ای انجام می‌دهند؛ به این ترتیب که به جای افزایش وضوح یک‌باره، فرایند افزایش وضوح به مراحل کوچک‌تری تقسیم می‌شود. این رویکرد باعث مدیریت بهتر و بهبود عملکرد در افزایش وضوح تصویر می‌شود.

۴) اتصالات بازگشتی[9]: نوع دیگری از معماری‌ها هستند که خروجی برخی لایه‌ها را به لایه‌های قبلی بازمی‌گردانند. این اتصالات، امکان استفاده چندباره از برخی لایه‌ها را فراهم می‌کنند و به بهبود عملکرد شبکه بدون افزایش تعداد لایه‌ها کمک می‌کنند.

۵) اتصالات مبتنی بر تمرکز[10]: این نوع از اتصالات به شبکه اجازه می‌دهند که بخش‌های مهم‌تر تصاویر را شناسایی کرده و آنها را با اولویت بیشتری پردازش کنند. این اتصالات باعث می‌شوند که شبکه به اطلاعات کلیدی بیشتری دست یابد و تصاویر باکیفیت‌تری تولید کند.

۶) اتصالات چندشاخه‌ای[11]: برای ادغام اطلاعات از چندین منبع و مقیاس به کار می‌روند. این نوع اتصالات به شبکه اجازه می‌دهند تا از اطلاعات مکمل و تکمیلی چندین لایه بهره ببرند و به این ترتیب وضوح و دقت بیشتری در پردازش و تولید تصاویر به دست آورد.

۷) اتصالات چگال متصل[12]: به منظور افزایش ارتباط بین لایه‌های مختلف به کار گرفته می‌شوند و هر لایه را به تمامی لایه‌های دیگر متصل می‌کنند تا مشکل محوشدگی گرادیان کاهش یابد.

۸) اتصالات مدیریت تخریب چندگانه[13]: در معماری‌هایی که با تخریب اطلاعات در تصاویر ورودی روبه‌رو هستند، از اتصالات مدیریت تخریب چندگانه استفاده می‌شود. این نوع اتصالات کمک می‌کنند تا شبکه با استفاده از اطلاعات چندین مقیاس یا سطح تخریب، میزان تخریب یا ازدست‌رفتگی اطلاعات را کاهش دهد و تصاویر باکیفیت‌تری تولید کند.

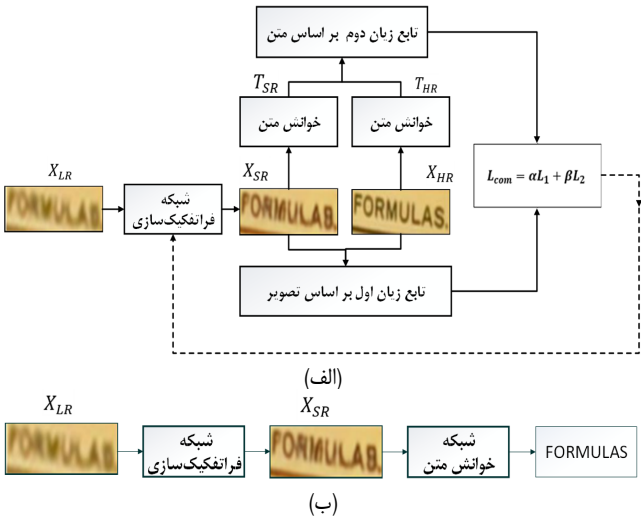
۹) شبکه‌های مولد مجادلانه[14]: به‌عنوان یکی از پیشرفته‌ترین معماری‌ها برای تولید تصاویر جدید و واقع‌گرایانه استفاده می‌شوند. این مدل‌ها شامل یک شبکه مولد و یک شبکه تمیزدهنده هستند. شبکه مولد تلاش می‌کند تا تصاویر مصنوعی تولید کند که به اندازه کافی واقع‌گرایانه به نظر برسند، در حالی که شبکه تمیزدهنده سعی دارد تفاوت بین تصاویر و تصاویر واقعی را تشخیص دهد. این رقابت موجب می‌شود که شبکه مولد نهایتاً بتواند تصاویر با تفکیک‌پذیری بالاتری تولید کند.

در جدول ۱، یک نمونه مطرح از هر دسته به همراه مراجع مرتبط با آن گزارش شده‌اند.

۳- روش پیشنهادی

در این تحقیق، هدف اصلی طراحی یک شبکه فراتفکیک‌سازی جدید و مختص تصاویر متنی برای بهبود کیفیت این نوع از تصاویر است. تصاویر متنی به دلیل پیچیدگی ساختاری حروف و اهمیت حفظ خوانایی، چالش‌هایی منحصربه‌فرد در فرایند فراتفکیک‌سازی به وجود می‌آورند که حل آنها نیازمند معماری‌های پیشرفته و رویکردهای نوین یادگیری است.

1. Linear Connection
2. Super-Resolution CNN
3. Up-Sampling
4. Fast SRCNN
5. Residual Connection
6. Very Deep SR
7. Residual Learning
8. Progressive Reconstruction Connection
9. Recursive Connection
10. Attention Based Connection
11. Multi Branch Connection
12. Dense Connected Connection
13. Multi Degradation Handling Connection
14. Generative Adversarial Network



شکل ۲: معماری پیشنهادی شبکه فراتفکیک‌سازی، (الف) مرحله آموزش و (ب) مرحله آزمایش.

یکی از عوامل مهم در طراحی شبکه‌های فراتفکیک‌سازی، تابع زیان است که نقش مستقیمی در هدایت وزن‌های شبکه در هنگام آموزش به سمت تولید تصویر با وضوح بالا دارد. در روش پیشنهادی، علاوه بر استفاده از توابع زیان کلاسیک که هدفشان شباهت هرچه بیشتر تصویر فراتفکیک‌سازی‌شده و تصویر اصلی است، از یک تابع زیان نوآورانه نیز بهره گرفته شده که بر حفظ خوانایی متن تمرکز دارد. این تابع زیان بر پایه یک شبکه خوانش متن طراحی شده که خروجی‌های متنی تصاویر اصلی و فراتفکیک‌سازی‌شده را با یکدیگر مقایسه می‌کند.

شکل ۲ شمای کلی روش پیشنهادی این مقاله را نشان می‌دهد. همان طور که در این شکل مشخص است، دو تابع زیان در این تحقیق پیشنهاد شده است. تابع زیان اول مسئولیت نزدیک‌شدن تصویر تولیدشده توسط شبکه به تصویر اصلی را بر عهده دارد. این تابع زیان در همه معماری‌های فراتفکیک‌سازی نقش اساسی دارد. تابع زیان دوم بر اساس متن خوانده‌شده توسط یک شبکه خوانش متن تنظیم می‌گردد. در نهایت حاصل جمع وزن‌دار این دو تابع زیان به‌عنوان فیدبک نهایی به شبکه فراتفکیک‌سازی داده می‌شود.

در این تحقیق برای تابع زیان اول از میانگین مربعات خطا[1] استفاده شده است

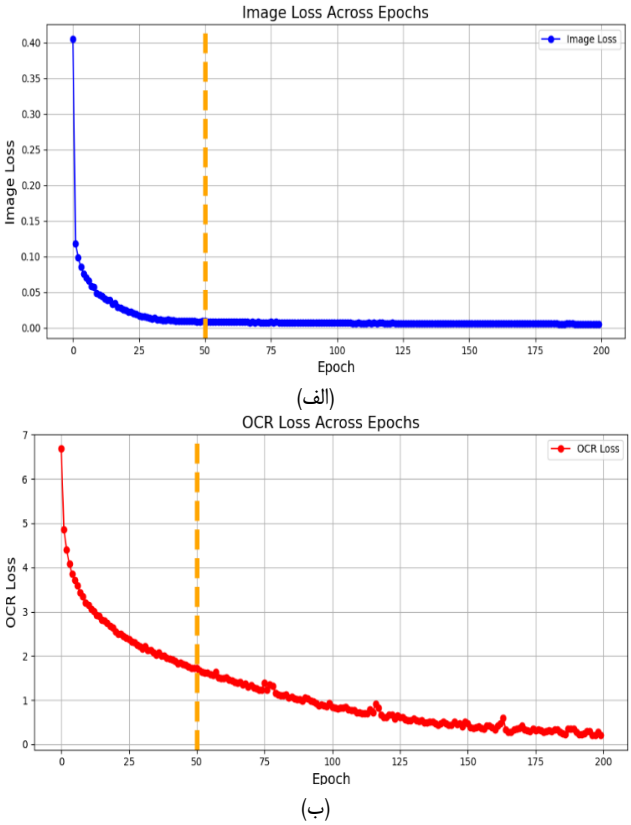
$$L_{image} = L_{MSE} = \frac{1}{N}\sum_{i=0}^{N}(I_{SR}(i) - I_{HR}(i))^2 \qquad (1)$$

که در آن N تعداد کل پیکسل‌ها در تصویر، $I_{SR}(i)$ مقدار پیکسل iام در تصویر SR و $I_{HR}(i)$ مقدار پیکسل iام در تصویر SR است. تابع زیان دوم که بر اساس متن خوانده‌شده تعریف می‌گردد به شکل زیر پیشنهاد شده است

$$L_{text} = \frac{1}{L}\sum_{j=1}^{L} L_{CE}(T_{SR}^{(j)}, T_{GT}^{(j)}) \qquad (2)$$

که در آن L تعداد حروف متن، T_{SR}^{j} بردار احتمال حرف jام در متن تشخیص‌داده‌شده در تصویر SR و $T_{GT}^{(j)}$ بردار وان‌هات[2] حرف jام برای متن واقعی تصویر HR است. منظور از بردار وان‌هات برداری است که همه درایه‌های آن به‌جز یک درایه صفر و مابقی درایه‌ها یک است.

1. Mean Squared Error
2. One-Hot



شکل ۳: نمودار زیان بر حسب تعداد دوره آموزشی، (الف) تابع زیان اول بر اساس تصویر و (ب) تابع زیان دوم بر اساس متن.

همچنین L_{CE} که در واقع آنتروپی متقاطع[3] [۲۷] بردار متن تشخیص داده داده شده و متن واقعی است به صورت زیر تعریف می‌شود

$$L_{CE}(T_{SR}^{j}, T_{SR}^{j}) = -\sum_{k=1}^{C} T_{GT}^{(j)}(k)\log(T_{SR}^{(j)}(k)) \qquad (3)$$

که C تعداد کلاس‌های الفبای مورد استفاده شامل تمامی حروف، اعداد و نمادهاست. واضح است که اگر بردار T_{SR} کاملاً مشابه T_{GT} باشد بدین معناست که شبکه خوانش متن با اطمینان ۱۰۰ درصدی متن را درست تشخیص داده که در این صورت مقدار تابع زیان صفر است. این تابع زیان تضمین می‌کند متن استخراج‌شده نه‌تنها باید با متن واقعی یکسان باشد، بلکه با اطمینان بالایی این تطابق را برقرار کند. استفاده از آنتروپی متقاطع به عنوان معیار مقایسه، شبکه را به سمت تخصیص بیشترین احتمال به حروف صحیح هدایت می‌کند که از نظر ریاضی به کاهش خطای پیش‌بینی منجر می‌شود. این امر موجب می‌گردد خطاهای احتمالی تا حد ممکن کاهش یافته و بازسازی تصویر متنی با دقت و اطمینان بالاتری انجام شود. نهایتاً زیان نهایی بدین صورت محاسبه می‌گردد

$$L_{total} = \alpha L_{image} + \beta L_{text} \qquad (4)$$

در مراحل اولیه آموزش، شبکه نیاز دارد تمرکز بیشتری بر حفظ ساختار تصویر داشته باشد تا فضای جستجو به فضای تصویر اصلی نزدیک‌تر شود. علاوه بر این همان طور که در شکل ۳ مشخص است، مقدار زیان مربوط به خوانش متن به‌طور محسوسی بیشتر از زیان بازسازی تصویر است. این عدم توازن موجب می‌شود شبکه به بهبود کیفیت تصویر بی‌توجه بوده و فرایند یادگیری به سمت بازسازی تصویر اصلی همگرا نشود؛ بنابراین $\alpha = 0.999$ و $\beta = 0.001$ یک انتخاب مناسب است.

3. Cross Entropy

شکل ۴: نمونه‌هایی از پایگاه داده TextZoom، (الف) نمونه‌های آسان برای خوانش، (ب) نمونه‌های معمولی برای خوانش و (ج) نمونه‌های سخت برای خوانش.

شبکه خوانش متن روش پیشنهادی، نقشی اساسی در تابع زیان ایفا می‌کند؛ بنابراین این شبکه باید به‌گونه‌ای طراحی شود که حساسیت و سخت‌گیری لازم را در مواجهه با تصاویر با وضوح پایین داشته باشد. اگر شبکه خوانش متن، توانایی بالایی در خواندن تصاویر با کیفیت پایین داشته باشد، ممکن است بازخورد دقیقی به شبکه اصلی ندهد.

برای روشن‌ترشدن این موضوع می‌توان آن را با وضعیت یک کودک باهوش با بینایی ضعیف مقایسه کرد که تلاش می‌کند تا علائم بینایی‌سنجی را حدس بزند. در این حالت، کودک ممکن است گاهی به طور تصادفی درست حدس بزند، اما مشکل اصلی بینایی او همچنان نادیده گرفته می‌شود و تشخیص صحیح انجام نمی‌شود. در واقع این کودک باهوش با ارائه بازخورد اشتباه به اپتومتریست، مانع از تشخیص درست مشکل می‌شود.

برای دستیابی به بازخورد معتبر و کمک به تولید تصاویر باکیفیت‌تر، تنظیم قدرت تشخیص شبکه خوانش متن ضروری است. این شبکه باید به‌گونه‌ای عمل کند که به صورت تنظیم‌شده، عملکرد شبکه اصلی را بهبود بخشد. به این رویکرد «تابع زیان با تضعیف عامدانه خوانش» می‌گوییم، چرا که با اعمال سخت‌گیری بر شبکه اصلی، آن را به سمت تولید تصاویر با تفکیک‌پذیری بالاتر و خواناتر هدایت می‌کند.

تنظیم دقیق و متعادل قدرت تشخیص شبکه خوانش متن، یکی از عوامل کلیدی در بهبود کیفیت تصاویر فراتفکیک‌شده در روش پیشنهادی است. برای ارضای این شرایط با الهام از [۲۸] که یک چهارچوب چهارمرحله‌ای برای خوانش متن از صحنه (STR)[1] ارائه می‌دهد، استفاده کردیم. این مراحل شامل دریافت تصویر ورودی، استخراج ویژگی‌های تصویری، مدل‌سازی توالی برای یادگیری وابستگی‌های متنی و پیش‌بینی کاراکترهاست. این مقاله با معرفی این چهارچوب چهارمرحله‌ای، امکان ارزیابی دقیق تأثیر ماژول‌های مختلف بر دقت، سرعت و مصرف حافظه را فراهم می‌کند. با حذف عامدانه BLSTM[2] از مرحله سوم، وابستگی مدل به پیش‌بینی‌های زنجیره‌ای کاهش داده شد تا صرفاً بر اطلاعات بصری تکیه کند. با اعمال این تغییر، به‌طور عامدانه توانایی مدل در جبران نواقص تصویر از طریق وابستگی‌های متنی تضعیف شد. این تضعیف عامدانه باعث سخت‌گیری بیشتر مدل در مواجهه با تصاویر کم‌کیفیت شد و در نتیجه، شبکه SR را به تولید خروجی‌های دقیق‌تر و شفاف‌تر وادار کرد. همچنین این شبکه نسبت به شبکه اصلی از تعداد لایه‌های کمتری تشکیل گردیده و به صورت آگاهانه صرفاً با داده‌های با تفکیک‌پذیری بالا آموزش داده شد.

٤- پایگاه داده

در این تحقیق از مجموعه داده‌های TextZoom [۲۹]، ICDAR [۳۰] و SVT [۳۱] برای آموزش و ارزیابی کارایی روش پیشنهادی استفاده شده است. نمونه‌هایی از تصاویر مجموعه داده TextZoom در شکل ۴ آمده است. این مجموعه داده شامل ۲۱۷۴۰ جفت تصویر با وضوح پایین و وضوح بالاست که از طریق زوم لنز دوربین در شرایط واقعی جمع‌آوری شده‌اند. مجموعه آموزشی شامل ۱۷۳۶۷ جفت تصویر بوده و مجموعه آزمایش بر اساس فاصله کانونی لنز دوربین که تأثیر مستقیمی بر کیفیت تصاویر ثبت‌شده دارد، به سه زیرمجموعه آسان (۱۶۱۹ نمونه)، متوسط (۱۴۱۱ نمونه) و سخت (۱۳۴۳ نمونه) تقسیم شده است.

مجموعه داده ICDAR۲۰۱۵، یکی از معروف‌ترین مجموعه داده‌های خوانش متن در صحنه، شامل ۲۰۷۷ تصویر برش‌خورده از متون موجود در عکس‌های خیابانی برای ارزیابی است. این تصاویر به دلیل ثبت تصادفی در خیابان، دارای وضوح پایین و تاری هستند که خوانش متن در آنها را چالش‌برانگیز می‌کند.

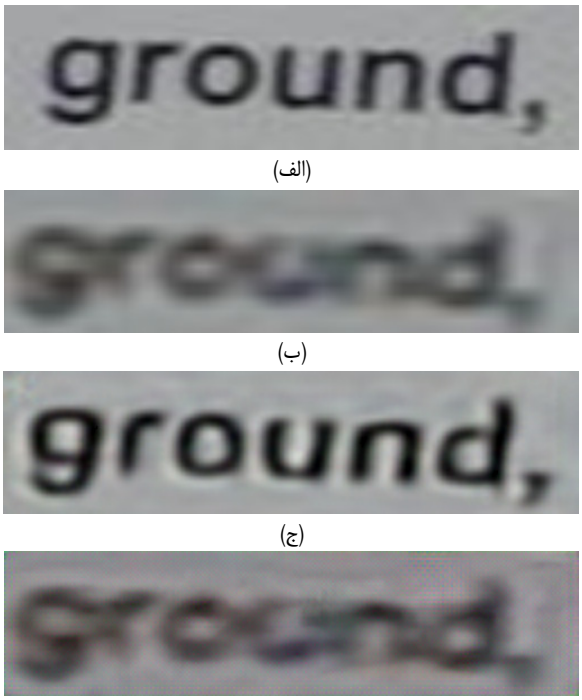
مجموعه داده SVT نیز یک مجموعه داده شناخته‌شده برای خوانش متن در صحنه است که شامل ۶۴۷ تصویر آزمایش است. به دلیل کیفیت پایین تصاویر گرفته‌شده در خیابان، خوانش متن در این مجموعه داده نیز چالش‌هایی به همراه دارد.

۵- شبیه‌سازی و تحلیل نتایج

در این بخش نشان داده شده که چگونه روش پیشنهادی منجر به تقویت یکی از معروف‌ترین معماری‌های شبکه‌های فراتفکیک‌سازی [۲۵] شده است. این تقویت از طریق بهبود بازخورد شبکه خوانش متن و تنظیم دقیق تابع زیان مبتنی بر تضعیف عامدانه به دست آمده است. به این ترتیب، کیفیت تصاویر فراتفکیک‌شده به طور قابل توجهی افزایش یافته و خوانایی متن در این تصاویر بهبود پیدا کرده که نشان‌دهنده اثربخشی و کارایی بالای روش پیشنهادی است (شکل ۵).

شکل ۳ نشان‌دهنده مرحله آموزش شبکه فراتفکیک‌سازی روش پیشنهادی می‌باشد. در این شکل، محور افقی هر دو نمودار نشان‌دهنده دوره‌های آموزشی و محور عمودی مقدار تابع زیان می‌باشد. شکل ۳- الف مربوط به تابع زیان شبکه بر اساس تصویر و شکل ۳- ب، تابع زیان بر اساس متن خوانده‌شده توسط شبکه خوانش متن است. مشاهده می‌شود که مقدار تابع زیان مربوط به تصویر پس از ۵۰ دور آموزشی، کاهش محسوسی نداشته و تقریباً به یک مقدار ثابت نزدیک شده است. این موضوع بیانگر این است که شبکه در بهبود کیفیت بصری تصویر به سطح مطلوبی رسیده و کاهش بیشتری در این معیار به دست نیامده است. از سوی دیگر، تابع زیان بر اساس خوانش متن، کاهش مداومی را از دور آموزشی ۵۰ به بعد نشان می‌دهد. این کاهش مداوم حاکی از آن است که شبکه فراتر از بهبود کیفیت بصری در بهبود خوانایی متن در تصاویر نیز موفق بوده است؛ بنابراین استفاده از خطای خوانش متن به عنوان تابع

1. Scene Text Recognition
2. Bidirectional Long Short-Term Memory



(الف) (ب) (ج) (د)

شکل ۵: یک نمونه از نتایج معماری پیشنهادی، (الف) تصویر HR، (ب) تصویر LR (ورودی شبکه پیشنهادی)، (ج) تصویر SR (خروجی شبکه پیشنهادی) و (د) تصویر SR (خروجی شبکه [۲۵]).

زیان و بازخورد آن، باعث هدایت شبکه به سمتی شده که نه‌تنها تصویر با وضوح بالاتری تولید کند، خوانایی متن موجود در آن افزایش یابد.

۵-۱ تحلیل ریاضی نتایج

تحلیل نتایج شکل ۳ نشان می‌دهد که پس از دوره‌های اولیه آموزش، کاهش خطای خوانش متن به میزان قابل توجهی سریع‌تر از کاهش خطای تصویر پیش می‌رود؛ بنابراین در بین تصاویر با خطای میانگین مربعات مشابه، زیرمجموعه‌ای از تصاویر وجود دارد که برای شبکه‌های خوانش متن، راحت‌تر قابل تشخیص هستند. دستیابی به این زیرمجموعه فقط با داشتن تابع زیان مبتنی بر تصویر شکل ۳- الف (i,j,k)، ممکن نیست. تحلیل زیر نشان می‌دهد که چگونه استفاده از تابع زیان پیشنهادی این تحقیق منجر به یافتن یک زیرمجموعه هدفمند از تصاویر شده است.

اگر تعداد کل پیکسل‌های تصویر خروجی $c\times m\times n$ باشد که در آن c تعداد کانال‌های رنگی، m تعداد ردیف‌ها و n تعداد ستون‌های تصویر باشد، در این صورت می‌توان فضای تمامی تصاویر ممکن خروجی را به صورت مجموعه S در نظر گرفت،

$$S=\{X\in R^{c\times m\times n}\,\big|\,\cdot\le X_{ijk}\le ۱\ ,\ \forall i,j,k\} \qquad (۵)$$

که در آن X_{ijk} مقدار به‌هنجارشده پیکسل در محل (i,j,k) با مقدار بین ۰ و ۱ می‌باشد. مشخص است که فضای جستجو در ابتدا بسیار بزرگ است و هدف آموزش رسیدن به بهترین زیرمجموعه از این زیرفضاست؛ به طوری که تصویر تولیدشده نهایی از نظر کیفیت تصویر و دقت خوانش، بیشترین شباهت را به تصویر واقعی داشته باشد.

پس از دوره آموزشی ۵۰ام، خطای میانگین مربعات بین تصویر SR و HR به یک حد آستانه $\varepsilon=۰/۰۰۹$ می‌رسد و ادامه آموزش منجر به کاهش معنادار میانگین مربعات نمی‌شود. در واقع جستجو برای یافتن بهترین تصویر از فضای بسیار بزرگ S به یک زیرمجموعه بسیار کوچک‌تر از خودش رسیده است به طوری که

ردیف	تصویر LR	تصویر SR	تصویر HR
۱	Reasternat Conf = 0.67	Restaurant Conf = 0.7432	Restaurant Conf = 0.9991
۲	While Conf = 0.7075	While Conf = 0.9995	While Conf = 0.9999
۳	actwity Conf = 0.6107	activity Conf = 0.8068	activity Conf = 0.9986
۴	not Conf = 0.7179	FACE Conf = 0.5319	FACE Conf = 0.5635
۵	11:30:2.30 Conf = 0.1859	11:30.2:30 Conf = 0.3240	11:30-2:30 Conf = 0.9739
۶	gm Conf = 0.1828	grass Conf = 0.7844	grass Conf = 0.9970
۷	Organiza Conf = 0.4743	Oeganic Conf = 0.7443	Organic Conf = 0.9999
۸	Cucpuscing Conf = 0.0074	Cappuccina Conf = 0.2653	Cappuccino Conf = 0.9976
۹	muthemuld Conf = 0.16	mathematics Conf = 0.2347	mathematics Conf = 0.9048
۱۰	Ansinguan Conf = 0.0108	Insingram Conf = 0.1554	Instagram Conf = 0.9980

شکل ۶: نمونه‌هایی از بهبود تصاویر متنی توسط شبکه پیشنهادی.

$$S_{MSE}\subset S \qquad (۶)$$

$$S_{MSE}=\{I_{SR}\in S\,\big|\,MSE(I_{SR},I_{HR})<\varepsilon\} \qquad (۷)$$

همان طور که در شکل ۳ مشخص است در زیرفضای S_{MSE} با پیش‌بردن آموزش، خطای میانگین مربعات کاهش چندانی نمی‌یابد؛ بنابراین بازخورد این خطا به شبکه به‌تنهایی قادر به بهینه‌سازی عملکرد شبکه نیست. استفاده از رویکرد پیشنهادی در این تحقیق باعث شده آموزش شبکه به شکل معناداری ادامه پیدا کند؛ به طوری که جستجو در زیرفضای S_{MSE} به‌صورت هدف‌داری به یک زیرمجموعه کوچک‌تر می‌رسد، یعنی

$$S_{OCR}=\{I_{SR}\in S_{MSE}\,\big|\,OCR(I_{SR})>\tau_{OCR}\} \qquad (۸)$$

$$S_{OCR}\subset S_{MSE}\subset S \qquad (۹)$$

که τ_{OCR} نشان‌دهنده یک آستانه برای دقت خوانش متن استخراج‌شده از تصویر SR است. رسیدن به زیرفضای S_{OCR} از دستاوردهای استفاده از تابع زیان معرفی‌شده در این تحقیق است.

۵-۲ تحلیل عملیاتی نتایج

در شکل ۶ ده نمونه از تصاویر با وضوح پایین موجود در مجموعه داده‌های آزمایشی، انتخاب شده تا عملکرد شبکه پیشنهادی را نشان دهد.

جدول ۲: دقت خوانش متن بر حسب درصد بر روی تصاویر پایگاه داده TEXTZOOM.

روش‌ها	آسان	متوسط	سخت
Bicubic	۶۱٫۳۹	۳۷٫۹۴	۲۰٫۶۲
SRResnet [۲۵]	۶۲٫۴۴	۳۹٫۲۲	۲۲٫۴۸
PAN [۳۲]	۶۲٫۶۳	۳۹٫۳۶	۲۲٫۱۱
DRLN [۳۳]	۶۲٫۳۸	۳۸٫۶۵	۲۲٫۰۴
A2N [۳۴]	۶۲٫۰۱	۳۹٫۲۲	۲۲٫۲۶
HAT [۳۵]	۶۲٫۱۹	۳۹٫۰۸	۲۲٫۴۱
DAT [۳۶]	۶۲٫۵۰	۳۸٫۸۶	۲۲٫۳۸
روش پیشنهادی	**۶۸٫۴۳**	**۴۷٫۵۲**	**۳۱٫۴۷**

در این شکل ستون اول، تصاویر ورودی با وضوح پایین (LR) را نمایش می‌دهد. این تصاویر به دلیل کیفیت پایین، هم از نظر بصری و هم از نظر خوانایی، کیفیت مطلوبی ندارند و متن موجود در آنها حتی برای انسان نیز قابل تشخیص نیست. ستون دوم، تصاویر فراتفکیک‌سازی‌شده (SR) را نشان می‌دهد که خروجی روش پیشنهادی هستند. این تصاویر نسبت به ورودی‌های LR از وضوح بالاتر و خوانایی بهتری برخوردارند، به‌گونه‌ای که متن آنها قابل تشخیص شده و کیفیت بصری آنها به طور محسوسی افزایش یافته است. ستون سوم، تصاویر مرجع با وضوح بالا (HR) را نمایش می‌دهد که به‌عنوان کیفیت ایده‌آل متن و ساختار حروف برای ارزیابی عملکرد روش پیشنهادی به کار رفته‌اند. هر ردیف نمایانگر یک نمونه از تصاویر پایگاه داده با سه کیفیت مختلف (منطبق بر هر ستون)، متن تشخیص‌داده‌شده توسط شبکه خوانش متن و ضریب اطمینان[۱] (Conf) مربوط به هر خوانش است. همان طور که مشاهده می‌شود، اعمال فراتفکیک‌سازی توسط روش پیشنهادی منجر به افزایش قابل توجه وضوح و خوانایی متن شده است، به‌گونه‌ای که سیستم خوانش متن با اطمینان بالاتر و دقت بیشتری قادر به خوانش محتوای متنی تصاویر بوده است.

در جدول ۲ دقت خوانش متن بر روی داده‌های آزمایش پایگاه داده TextZoom گزارش شده است. همان طور که دیده می‌شود، افزودن تابع زیان معرفی‌شده در این تحقیق به معماری [۲۵] در قالب روش پیشنهادی، منجر به بهبود کیفیت تصاویر و افزایش دقت خوانش متن در مقایسه با نسخه اصلی شده است.

۶- نتیجه‌گیری

در این پژوهش، یک روش نوآورانه برای فراتفکیک‌سازی تصاویر متنی ارائه شد که به‌طور ویژه بر حفظ ساختار حروف و ارتقای خوانایی متن تمرکز دارد. روش پیشنهادی با بهره‌گیری از ترکیب دو تابع زیان، شامل زیان مبتنی بر شباهت تصویری و زیان خوانش متن، توانسته است کیفیت تصاویر متنی را به شکل قابل توجهی بهبود بخشد.

نتایج تجربی نشان می‌دهد این رویکرد نه‌تنها وضوح بصری تصاویر را افزایش می‌دهد، بلکه دقت سیستم‌های خوانش متن را به‌طور چشمگیری بهبود می‌بخشد. استفاده از بازخورد سخت‌گیرانه شبکه خوانش از طریق تضعیف عامدانه آن، امکان بازیابی اطلاعات ازدست‌رفته در تصاویر با وضوح پایین را فراهم کرده و به تولید تصاویر فراتفکیک‌سازی‌شده با جزئیات بیشتر و خوانایی بالاتر منجر می‌شود.

۱. میانگین بیشینه احتمال‌های بردارهای خروجی شبکه برای هر کاراکتر در متن است که نشان‌دهنده میزان اطمینان شبکه در بازشناسی متن می‌باشد.

روش پیشنهادی با افزایش فراتفکیک‌پذیری تصاویر وضوح پایین، گامی مؤثر در بهبود عملکرد سیستم‌های OCR و کاربردهای مرتبط محسوب می‌شود. علاوه بر این، بهینه‌سازی بیشتر این روش از طریق بهره‌گیری از معماری‌های جدیدتر و تلفیق با توابع زیان دیگر می‌تواند به‌عنوان یک مسیر پژوهشی مورد بررسی قرار گیرد.

مراجع

[1] R. Shu, C. Zhao, S. Feng, L. Zhu, and D. Miao, "Text-enhanced scene image super-resolution via stroke mask and orthogonal attention," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6317-6330, Nov. 2023.

[2] J. Ma, S. Guo, and L. Zhang, "Text prior guided scene text image super-resolution," *IEEE Trans. on Image Processing*, vol. 32, pp. 1341-1353, 2023.

[3] J. Ma, Z. Liang, and L. Zhang, "A text attention network for spatial deformation robust scene text image super-resolution," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 5911-5920, New Orleans, LA, USA, 19-24 Jun. 2022.

[۴] ع. عابدی و ا. کبیر، "فراتفکیک‌پذیری مبتنی بر نمونه تک‌تصویر متن با روش نزول گرادیان ناهمزمان ترتیبی،" *نشریه مهندسی برق و مهندسی کامپیوتر ایران، ب- مهندسی کامپیوتر*، سال ۱۴، شماره ۳، صص. ۱۹۲-۱۷۷، پاییز ۱۳۹۵.

[5] K. Mehrgan, A. R. Ahmadyfard, and H. Khosravi, "Super-resolution of license-plates using weighted interpolation of neighboring pixels from video frames," *International J. of Engineering, Trans. B: Applications*, vol. 33, no. 5, pp. 992-999, May 2020.

[6] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. 13th European Conf. Computer Vision,* pp. 184-199, Zurich, Switzerland, 6-12 Sept. 2014.

[7] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput Imaging*, vol. 2, no. 2, pp. 109-122, Jun. 2016.

[8] M. Hradiš, J. Kotera, P. Zemcık, and F. Šroubek, "Convolutional neural networks for direct text deblurring," in *Proc. of the British Machine Vision Conf.*, 13 pp., Swansea, UK, 7-10 Dec. 2015.

[9] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal Mach Intell*, vol. 38, no. 2, pp. 295-307, Feb. 2015.

[10] D. Gudivada and P. K. Rangarajan, "Enhancing PROBA-V satellite imagery for vegetation monitoring using FSRCNN-based super-resolution," in *Proc. Int. Conf. on Next Generation Electronics,* 6 pp., Vellore, India, 14-16 Dec. 2023.

[11] J. Zhang, M. Liu, X. Wang, and C. Cao, "Residual net use on FSRCNN for image super-resolution," in *Proc. 40th Chinese Control Conf.,* pp. 8077-8083, Shanghai, China, 26-28 Jul. 2021.

[12] T. Khachatryan, D. Galstyan, and E. Harutyunyan, "A comprehensive approach for enhancing deep learning datasets quality using combined SSIM algorithm and FSRCNN," in *Proc. IEEE East-West Design & Test Symp.*, 4 pp., 22-25 Sept. 2023.

[13] Y. Zhu, X. Sun, W. Diao, H. Li, and K. Fu, "RFA-Net: reconstructed feature alignment network for domain adaptation object detection in remote sensing imagery," *IEEE J. Sel Top Appl Earth Obs Remote Sens*, vol. 15, pp. 5689-5703, 2022.

[14] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 370-378, Santiago, Chile, 7-13 Dec. 2015.

[15] M. Chen, *et al.*, "RFA-Net: residual feature attention network for fine-grained image inpainting," *Engineering Applications of Artificial Intelligence*, vol. 119, Article ID: 105814, Mar. 2023.

[16] Z. Wang and J. Tang, "Advancing quality and detail: enhanced-lapSRN for chip socket image super-resolution," in *Proc. Int. Conf. on Image Processing, Computer Vision and Machine Learning*, pp. 153-159, Chengdu, China, 3-5 Nov. 2023.

[17] R. Tang, *et al.*, "Medical image super-resolution with Laplacian dense network," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3131-3144, Jan. 2022.

[18] K. Wu, C. K. Lee, and K. Ma, "Memsr: training memory-efficient lightweight model for image super-resolution," in *Proc. 39th Int. Conf. on Machine Learning,* pp. 24076-24092, Baltimore, MD, USA, 17-23 Jul. 2022.

[19] Z. Du, *et al.*, "Fast and memory-efficient network towards efficient image super-resolution," in *Proc. of the IEEE/CVF Conf. on*

Computer Vision and Pattern Recognition, pp. 853-862, New Orleans, LA, USA, 19-20 Jun. 2022.
[20] K. H. Liu, B. Y. Lin, and T. J. Liu, "MADnet: a multiple attention decoder network for segmentation of remote sensing images," in *Proc. Int. Conf. on Consumer Electronics-Taiwan* pp. 835-836, PingTung, Taiwan, 17-19 Jul. 2023.
[21] D. Zhang, W. Zhang, W. Lei, and X. Chen, "Diverse branch feature refinement network for efficient multi-scale super-resolution," *IET Image Process*, vol. 18, no. 6, pp. 1475-1490, May 2024.
[22] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 4799-4807, Venice, Italy, 22-29 Oct. 2017.
[23] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3262-3271, Salt Lake City, UT, USA, 18-22 Jun. 2018.
[24] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "Ranksrgan: super resolution generative adversarial networks with learning to rank," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 10, pp. 7149-7166, Oct. 2021.
[25] C. Ledig, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4681-4690, Honolulu, HI, USA, 21-26 Jul. 2017.
[26] B. K. Xie, S. B. Liu, and L. Li, "Large-scale microscope with improved resolution using SRGAN," *Optics & Laser Technology*, vol. 179, Article ID: 111291, Dec. 2024.
[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
[28] J. Baek, *et al.*, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 4715-4723, Seoul, South, Korea, 27 Oct.-2 Nov. 2019.
[29] W. Wang, *et al.*, "Scene text image super-resolution in the wild," in *Proc. 16th European Conf. on Computer Vision,* pp. 650-666, Glasgow, UK, 20-28 Aug. 2020.
[30] D. Karatzas, *et al.*, "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. on Document Analysis and Recognition*, pp. 1156-1160, Tunis, Tunisia, 23-26 Aug. 2015.
[31] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. Int. Conf. on Computer Vision.* pp. 1457-1464, Barcelona, Spain, 6-13 Nov. 2011.
[32] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Proc., Computer Vision-ECCV Workshops*, pp. 56-72, Glasgow, UK, 23-28 Aug. 2020.
[33] S. Anwar and N. Barnes, "Densely residual laplacian super-resolution," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 3, pp. 1192-1204, Mar. 2022.
[34] H. Chen, J. Gu, and Z. Zhang, *Attention in Attention Network for Image Super-Resolution*, arXiv Preprint, arXiv:2104.09497, 2021.
[35] X. Chen, X. Wang, J. Zhou, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 22367-22377, Vancouver, Canada, 18-22 Jun, 2023.
[36] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, "Dual aggregation transformer for image super-resolution," in *Proc. IEEE/CVF Int. Conf. on Computer Vision*, pp. 12278-12287, Vancouver, Canada, 18-22 Jun, 2023.

کمیل مهرگان تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی برق گرایش مخابرات سیستم به‌ترتیب در سال‌های ۱۳۹۶ و ۱۳۹۹ در دانشگاه صنعتی شاهرود به پایان رساند و از ۱۳۹۹ به دوره دکتری تخصصی مهندسی برق گرایش مخابرات سیستم در دانشگاه فردوسی مشهد وارد شد. زمینه‌های علمی مورد علاقه وی شامل موضوعاتی مانند پردازش سیگنال، پردازش تصاویر و یادگیری ماشین است.

عباس ابراهیمی مقدم مدرک کارشناسی و کارشناسی ارشد برق گرایش مخابرات خود را به‌ترتیب از دانشگاه‌های صنعتی شریف در سال ۱۳۷۰ و صنعتی خواجه نصیر طوسی در سال ۱۳۷۴ اخذ کرد. وی مدرک دکتری خود را از دانشگاه McMaster کانادا دریافت کرد و از سال ۱۳۹۰ به‌عنوان استادیار در دانشگاه فردوسی مشهد فعالیت علمی می‌نماید. زمینه‌های تحقیقاتی مورد علاقه وی پردازش گفتار، پردازش تصویر و ویدئو، بینایی ماشین و پردازش سیگنال‌های حیاتی است.

مرتضی خادمی درح تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی برق به‌ترتیب در سال‌های ۱۳۶۴ و ۱۳۶۶ در دانشگاه صنعتی اصفهان به پایان رساند. ایشان از سال ۱۳۶۶ تا ۱۳۷۰ به عنوان عضو هیأت علمی در دانشگاه فردوسی مشهد به کار مشغول بود. پس از آن به دوره دکترای مهندسی برق در دانشگاه Wollongong استرالیا وارد شد و در سال ۱۳۷۴ موفق به اخذ درجه دکترا در مهندسی برق از دانشگاه مذکور گردید. دکتر خادمی از سال ۱۳۷۴ مجدداً در دانشکده مهندسی دانشگاه فردوسی مشهد مشغول به فعالیت شد و اینک نیز استاد این دانشکده است. زمینه‌های علمی مورد علاقه وی شامل موضوعاتی مانند مخابرات ویدئویی، فشرده‌سازی ویدئو، پردازش تصویر و سیگنال‌های پزشکی و پنهان‌سازی اطلاعات در ویدئو است.