New Researches in the Smart City
Vol 3 (Issue 3), 6-18
nrsc.apadana.ac.ir

Apadana Institute of Higher Education

A Diffusion-Based Framework for Deep Learning Video Colorization

Seyyed Mehdi Hazratifard *
Chief Technology Officer, Soshianest Inc., Vancouver, Canada.

Amir Chekini
Assistant Professor, Department of Electrical and Computer Engineering, Faculty of Engineering, University of Victoria, Canada.

Received: 07/April/2025 Accepted: 31/May/2025 eISSN: 3060-6144 ISSN: 2980-8936

Abstract

Recent advancements in artificial intelligence have enabled the development of sophisticated algorithms capable of automatically colorizing black-and-white videos. This progress offers exciting opportunities to revitalize historical footage and provide content creators with powerful tools for visual storytelling. These techniques leverage deep neural networks to analyze and understand image content, detect visual patterns, and infer plausible color information—paving the way for richer semantic interpretation in computer vision. While most prior research has focused on image colorization, the application of deep learning techniques to video colorization remains relatively underexplored. In this work, we aim to bridge that gap by demonstrating that modern image colorization methods can be effectively adapted to video frames. Our solution aligns with current state-of-the-art approaches showcased at the NTIRE 2023 Video Colorization Challenge. Specifically, we investigate the use of diffusion models-powerful generative frameworks known for their success in image and text generation. In our implementation, noise is gradually introduced into video frames, and a U-Net architecture enhanced with self-attention mechanisms is employed to predict denoised (i.e., colorized) versions. The model is trained using the DAVIS and LDV datasets. Experimental evaluations comparing the predicted frames with ground truth across several quality metrics-including PSNR, SSIM, FID, and CDC-show promising results that validate the effectiveness of our approach.

Keywords: video colorization, deep learning diffusion models, evaluation metrics

* Corresponding Author: smhazrati@uvic.ca

فصلنامه پژوهش‌های نوین در شهر هوشمند
دوره ۳، شماره ۳، بهار ۱۴۰۴، ۶-۱۸
nrsc.apadana.ac.ir

مؤسسه آموزش عالی آپادانا
غیر دولتی - غیر انتفاعی

تاریخ دریافت: ۱۴۰۴/۰۱/۱۸ تاریخ پذیرش: ۱۴۰۴/۰۳/۱۰ eISSN: 3060-6144 ISSN: 2980-8936

رنگی‌سازی ویدئو بر پایه پیاده‌سازی مدل انتشار

سید مهدی حضرتی فرد* گروه علوم کامپیوتر، دانشگاه ویکتوریا، ویکتوریا، بریتیش کلمبیا، کانادا.

امیر چکینی گروه علوم کامپیوتر، دانشگاه ویکتوریا، ویکتوریا، بریتیش کلمبیا، کانادا.

چکیده

پژوهشگران در حال بهره‌گیری از تکنیک‌های پیشرفته برای توسعه الگوریتم‌هایی هستند که توانایی افزودن خودکار رنگ به ویدئوهای سیاه و سفید را دارند. این پیشرفت می‌تواند تجربه ما از فیلم‌های تاریخی را متحول کرده و ابزاری قدرتمند را در اختیار فیلم‌سازان و تولیدکنندگان ویدئو قرار دهد. این الگوریتم‌ها با استفاده از شبکه‌های عصبی عمیق پیشرفته به تحلیل تصاویر می‌پردازند، الگوها را شناسایی می‌کنند و راهی نویدبخش برای استخراج معنا و بینش از داده‌های بصری در حوزه بینایی ماشین ارائه می‌دهند. با وجود اینکه مطالعات کنونی، بیشتر بر رنگی‌سازی تصاویر متمرکز هستند، در زمینه ویدئوها و فیلم‌ها، همچنان یک خلأ محسوس در استفاده از تکنیک‌های یادگیری ماشین عمیق وجود دارد. این تحقیق با هدف پر کردن این شکاف صورت گرفته است و نشان می‌دهد که تکنیک‌های رنگی‌سازی تصاویر امروزی می‌توانند برای ویدئوها نیز به‌طور مؤثر استفاده شوند و با پیشرفته‌ترین روش‌های موجود معرفی‌شده در چالش رنگی‌سازی ویدئو NTIRE 2023، برابری کنند.

در این پژوهش، کاربرد مدل‌های انتشار مورد بررسی قرار گرفته است؛ مدل‌هایی که به دلیل توانایی در تولید تصویر و متن، محبوبیت یافته‌اند. در پیاده‌سازی ما، یک مدل انتشار برای افزودن نویز به فریم‌ها به کار رفته و یک شبکه U-Net مجهز به لایه‌های خودتوجهی، وظیفه پیش‌بینی فریم‌های بدون نویز و درنتیجه، پیش‌بینی رنگ فریم‌های ویدئویی را بر عهده دارد. برای آموزش مدل از مجموعه داده‌های DAVIS و LDV استفاده شد. بدین ترتیب، مقایسه فریم‌های رنگی‌شده با فریم‌های واقعی در مجموعه آزمون، نتایجی امیدوارکننده را در معیارهای کیفیت مختلف ازجمله PSNR، SSIM، FID و CDC نشان داد.

کلیدواژه‌ها: رنگی‌سازی ویدئو، مدل‌های انتشار در یادگیری عمیق، معیارهای ارزیابی

* نویسندهٔ مسئول: smhazrati@uvic.ca

۱- مقدمه

نقش رنگ در شکل‌گیری درک و فهم ما از دنیای بصری، به‌ویژه در زمینه ویدئو، انکارناپذیر است. بااین‌حال، تعداد زیادی از ویدئوهای قدیمی صرفاً به‌صورت سیاه و سفید وجود دارند که نوعی حس قدمت و گسست از دنیای معاصر را القا می‌کنند. پژوهشگران برای رفع این مسئله از مدل‌های یادگیری عمیق جهت بازیابی اطلاعات رنگی ازدست‌رفته در این ویدئوها استفاده کرده‌اند. این حوزه تحقیقاتی که با عنوان «رنگی‌سازی ویدئو با یادگیری عمیق»[1] شناخته می‌شود، در تلاش برای توسعه الگوریتم‌هایی است که قادر به افزودن خودکار رنگ به ویدئوهای سیاه و سفید باشند.

علاوه بر این، چالش «روندهای نوین در بازسازی و بهبود تصویر» (NTIRE) که توسط بنیاد چشم‌انداز کامپیوتر[2] پیشنهاد شده، فرصتی را برای پیشبرد این حوزه پژوهشی در میان جامعه محاسبات تصویری فراهم کرده است (Kang et al., 2023). این چالش از محققان دعوت می‌کند تا راه‌حل‌هایی را برای مسئله رنگی‌سازی ویدئو ارائه دهند.

ادبیات موجود در زمینه تولید تصویر، نشان‌دهنده علاقه گسترده به مدل‌های انتشار است (Dhariwal & Nichol, 2021)؛ همان‌گونه که در سایر حوزه‌های چشم‌انداز کامپیوتر نیز دیده می‌شود (Xu et al., 2023). این مدل‌ها با الهام از مفاهیم ترمودینامیکی طراحی شده و نقشی مشابه با نورون‌های اولیه در شبکه‌های عصبی را در شکل‌گیری پیشرفت‌های اخیر در یادگیری ماشین ایفا کرده‌اند. با توجه به این تحول در معماری‌های تولید داده، هدف مقاله ما، نمایش کاربرد مدل احتمالاتی انتشار عمیق[3] در تولید کانال‌های رنگی گمشده در فریم‌های ویدئویی است.

۱-۱- رنگی‌سازی

در ادبیات موجود پیرامون رنگی‌سازی ویدئو با یادگیری عمیق (DLVC)، سه رویکرد رایج برای افزودن رنگ به فریم‌های ویدئویی استفاده می‌شود: مبتنی بر خط‌خطی[4]، مبتنی بر نمونه[5] و کاملاً خودکار[6] (Stival & Pedrini, 2023).

در بین این روش‌ها، روش مبتنی بر خط‌خطی به‌عنوان یک رویکرد سنتی شناخته می‌شود که قدمت آن به دوران پیش از رواج یادگیری عمیق بازمی‌گردد. این روش، عمدتاً با انتقال رنگ از پیکسل‌های مجاور با توجه به شباهت در مقادیر روشنایی عمل می‌کند؛ اما فاکتورهایی نظیر بافت و زمینه اشیاء را نادیده می‌گیرد. به همین دلیل، امروزه کاربرد این رویکرد کمتر شده است؛ چراکه نیازمند دخالت گسترده انسانی است. با وجود این، برخی پژوهش‌ها همچنان از این رویکرد در وظایف DLVC استفاده کرده‌اند (Dogan et al., 2015; Heu et al., 2009; Yatziv & Sapiro, 2006).

رویکرد مبتنی بر نمونه، هم‌زمان با رشد یادگیری ماشین محبوب شد. در این روش، رنگی‌سازی با استفاده از نمونه‌هایی از تصاویر یا ویدئوهای مرجع انجام می‌شود، مانند طراحی خطی (Shi et al., 2020; Zhang et al., 2019)، بازسازی ویدئو (Meyer et al., 2018; Iizuka & Simo-Serra, 2019)، جریان اشیاء (Huang et al., 2022) و وابستگی‌های فضایی-زمانی (Chen et al., 2023). برخی دستاوردهای مهم در این حوزه شامل مدل‌های BisTNet، ColorVid و DeOldify هستند (Zhang et al., 2019; Wan et al., 2020; Salmona et al., 2022).

1. Deep learning video colorization (DLVC)
2. Computer vision foundation (CVF)
3. Deep diffusion probabilistic model (DDPM)
4. scribble-based
5. example-based
6. fully automatic

در مقایسه با روش خط‌خطی، ایجاد تطابق میان نمونه و فریم هدف در روش مبتنی بر نمونه پیچیده‌تر است. اغلب، این روش تلاش می‌کند همبستگی بین رنگ نمونه ورودی و روشنایی فریم هدف را یاد بگیرد؛ بنابراین، توانایی تعمیم این رویکرد به‌شدت تحت تأثیر تنوع داده‌های آموزشی است.

در میان این سه رویکرد، روش کاملاً خودکار دارای جایگاه برجسته‌ای است. برخلاف دو روش دیگر، این رویکرد در فرایند پیش‌بینی رنگ به نمونه‌های مرجع وابسته نیست و درنتیجه، در بسیاری از مدل‌های مدرن DLVC به کار گرفته شده است (Kouzouglidis et al., 2019; Lei & Chen, 2019; Liu et al., 2021; Mahajan et al., 2021). این روش‌ها از تکنیک‌هایی مانند GAN ساده، جریان زمانی، گسترش خودتنظیم و انتقال رنگ فریم کلیدی استفاده می‌کنند.

ویژگی مهم این رویکرد، تمرکز آن بر درک اشیاء موجود در تصویر، نورپردازی و ویژگی‌های زمانی برای حفظ یکنواختی رنگ است. این عوامل، رنگی‌سازی بدون نمونه را به روشی پیچیده‌تر نسبت به دیگر رویکردها تبدیل می‌کنند.

در پیاده‌سازی ما از رویکرد کاملاً خودکار استفاده شده است؛ چراکه این روش امکان کنترل بیشتر بر خروجی رنگی را فراهم می‌کند. بدین ترتیب، توسعه روشی که بتواند اطلاعات اشیاء موجود در فریم و زمینه آن‌ها را به مدل انتشار منتقل کند، برای ایجاد رنگ دقیق ضروری است.

۱-۲- مشارکت‌ها

تمرکز اصلی این پژوهش، طراحی رویکردی مبتنی بر مدل انتشار به‌منظور تولید رنگ در فریم‌های سیاه و سفید ویدئو است؛ به‌گونه‌ای که نتایج نهایی، بیشترین شباهت را با نسخه‌های رنگی اصلی داشته باشند. مشارکت‌های اصلی این مقاله را می‌توان به‌صورت زیر خلاصه کرد:

۱. پیاده‌سازی الگوریتمی که قادر است با استفاده از تکنیک انتشار شرطی، فریم‌های سیاه و سفید ویدئو را رنگی‌سازی کند.
۲. نمایش توانایی مدل در تولید نتایج با کیفیت بالا برای ویدئوهایی که در داده‌های آموزشی حضور نداشته‌اند که نشان‌دهنده رویکردی عمومی و قابل تعمیم است.
۳. کمک به پیشرفت مدل‌های انتشار در مسائل رنگی‌سازی ویدئو با یادگیری عمیق (DLVC) از طریق معرفی بهبودهای معماری نوآورانه.

۲- مدل‌های انتشار

مدل‌های انتشار، ازجمله شبکه‌های تخاصمی مولد (GANs) (Goodfellow et al., 2020)، خودرمزگذارهای واریاسیونی (VAEs) (Kingma & Welling, 2013) و برخی مدل‌های خودرگرسیو (Ramesh et al., 2021)، زیرمجموعه‌ای از مدل‌های مولد محسوب می‌شوند. در سال‌های اخیر، استفاده از این مدل‌ها رشد قابل‌توجهی داشته و باعث ارتقای کیفیت خروجی‌ها و تنوع روش‌های تولید تصویر شده است (Croitoru et al., 2023).

مدل غالب در حوزه بینایی ماشین که تقریباً در تمام مسائل تولید تصویر مورد استفاده قرار می‌گیرد، مدل احتمالاتی انتشار عمیق (DDPM) است (Sohl-Dickstein et al., 2015). این مدل‌ها که بر پایه اصول ترمودینامیک بنا شده‌اند، نتایج درخشانی را در بسیاری از وظایف بینایی ماشین ثبت کرده‌اند و به‌عنوان روش‌های پیشرفته حال حاضر شناخته می‌شوند (Dhariwal & Nichol, 2021).

فرایند مدل DDPM را می‌توان به‌صورت یک زنجیره مارکوف در نظر گرفت. در مرحله پیش رو، نویزهایی کوچک طی چند مرحله به تصویر افزوده می‌شوند. مرحله معکوس که بخش یادگیری‌پذیر شبکه است، تلاش می‌کند نویز افزوده‌شده را پیش‌بینی و حذف کند (Ho et al., 2020). در این مرحله، معمولاً از خودرمزگذارها برای یادگیری فرآیند حذف نویز استفاده می‌شود.

برای توضیح ساده، فرض کنید تصویری بدون نویز داریم که با x_0 نمایش داده می‌شود. در فرآیند انتشار، در هر مرحله t، مقداری نویز به آن افزوده شده و نمونه‌ها به x_1، x_2، ...، x_t تبدیل می‌شوند تا درنهایت به نویز خالص تبدیل گردند. در مسیر معکوس، مدل با یک نمونه نویزی تصادفی آغاز می‌کند که هدف آن، حذف تدریجی نویز و بازسازی x_0 است.

مدل‌های DDPM به‌طور گسترده در حوزه‌های مختلف محاسبات تصویری ازجمله وضوح‌بخشی (Kawar et al., 2022)، تولید تصویر (Preechakul et al., 2022)، ویرایش (Kim et al., 2022) و کاربردهای چندوجهی مانند ترمیم (Avrahami et al., 2022) و قطعه‌بندی (Rombach et al., 2022) به کار گرفته شده‌اند.

۳- کارهای پیشین

در این بخش، اهمیت موضوع رنگی‌سازی و کاربرد مدل‌های انتشار در آن به‌صورت جامع بررسی می‌شود. پیشرفت‌های چشمگیری در زمینه رنگی‌سازی تصویر با استفاده از فرآیندهای انتشار به ثبت رسیده‌اند. مطالعات مختلفی با پیاده‌سازی‌های متنوع پیشنهاد شده‌اند که هدف همه آن‌ها، دستیابی به نتایج طبیعی‌تر و واقعی‌تر است. این پیشرفت‌ها، مرزهای تکنیک‌های رنگی‌سازی را گسترش داده‌اند.

یکی از نوآوری‌های مهم، استفاده از روش‌های بدون ناظر برای نمونه‌برداری‌های پسین است که موجب افزایش سرعت آموزش و استنتاج شده و همچنین، قابلیت کاربرد در مسائلی چون ترمیم و کاهش تاری را فراهم کرده است (Kawar et al., 2022).

مدل GPIM[1]، نمونه‌ای دیگر است که عمدتاً برای رنگی‌سازی طراحی‌های خطی استفاده می‌شود. این مدل، امکان تولید چندین نسخه رنگی از یک ورودی را فراهم کرده و از جاسازی موقعیتی برای پر کردن دقیق‌تر رنگ‌ها بین خطوط بهره می‌برد (Furusawa et al., 2021).

مدل‌های انتشار در کاربردهای گسترده‌ای برای پردازش ویدئو مورد استفاده قرار گرفته‌اند، ازجمله ویرایش مبتنی بر متن با استفاده از فریم‌های کلیدی (Ceylan et al., 2023) و تولید ویدئو از توصیف‌های متنی (Ho et al., 2022; Yang et al., 2022; Zhou et al., 2022). در این رویکردها، متن به بردار جاسازی‌شده تبدیل گشته و مدل انتشار ویدئویی مطابق آن را تولید می‌کند.

در ادامه کاربردهای DDPM در ویدئو، رویکردهایی نیز در زمینه درون‌یابی فریم معرفی شده‌اند که فریم‌های میانی را برای حفظ یکنواختی زمانی تولید می‌کنند (Voleti et al., 2022).

با وجود حجم زیاد تحقیقات در زمینه تولید ویدئو با DDPM، در حوزه رنگی‌سازی ویدئو با یادگیری عمیق (DLVC) هنوز فاصله قابل‌توجهی از نظر کیفیت و کمیت پژوهش‌ها مشاهده می‌شود. این خلأ پژوهشی، همان چیزی است که این مطالعه در تلاش برای پر کردن آن است.

1. generative probabilistic image colorization

۴- روش‌شناسی

در این بخش، روش‌شناسی ما ارائه می‌شود؛ بدین صورت که ابتدا با معرفی داده‌های مورد استفاده و ویژگی‌های کلیدی آن‌ها آغاز می‌کنیم. سپس به تشریح معماری مدل می‌پردازیم و در پایان، جزئیات مربوط به فرآیند آموزش و ارزیابی مدل ارائه می‌شود.

۴-۱- مجموعه داده‌ها

برای آموزش مدل خود از دو مجموعه داده DAVIS (Voigtlaender et al., 2019) و LDV (Yang & Timofte, 2021) استفاده کردیم. مجموعه داده DAVIS شامل ۱۲۰ ویدئو است که به مجموعه‌های آموزش و اعتبارسنجی تقسیم شده‌اند؛ درحالی‌که مجموعه LDV شامل ۲۰۰ ویدئو برای آموزش است.

پیش از ورود تصاویر به مدل، آن‌ها به ابعاد ۲۲۴×۲۲۴ پیکسل تغییر اندازه داده شدند. همچنین از تکنیک‌های افزایش داده شامل برش تصادفی و چرخش برای بهبود توانایی تعمیم مدل استفاده کردیم.

در چارچوب این پژوهش، داده‌ها به سه بخش تقسیم شدند: فریم‌های سیاه و سفید ($Sg \in R^{1\times H\times W}$)، فریم‌های اصلی رنگی ($R \in R^{3\times H\times W}$) و فریم‌های رنگی تولیدشده توسط مدل ما ($Sc \in R^{3\times H\times W}$) که در آن، H و W ارتفاع و عرض فریم‌ها هستند.

۴-۲- مدل

معماری ما از چهار مؤلفه تشکیل شده است که مدل انتشار، هسته اصلی آن را تشکیل می‌دهد. در ادامه، اجزاء مختلف مدل و نحوه تعامل آن‌ها با یکدیگر تشریح می‌شود.

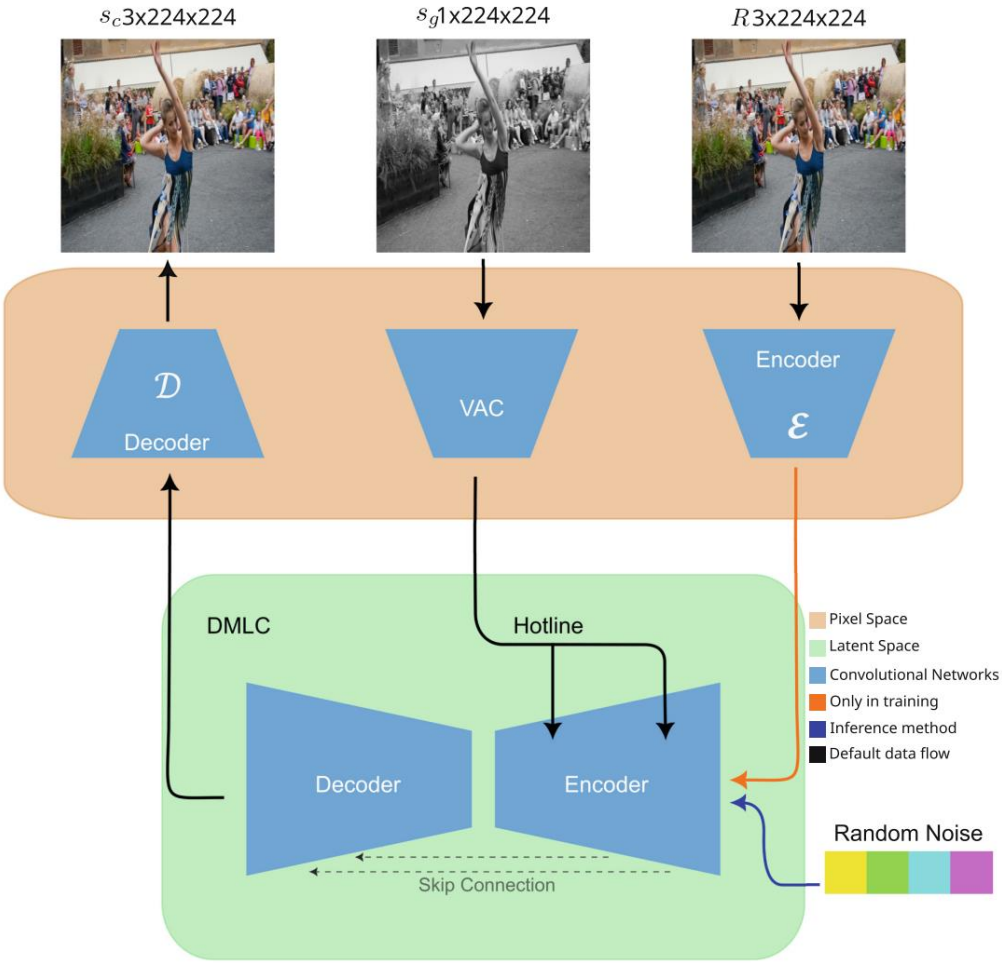
Encoder: رمزگذار (Encoder) با هدف استخراج فضای نهان[1] از فریم اصلی رنگی R طراحی شده است. خروجی آن، $LatR \in R^{4\times 28\times 28}$ است و به‌عنوان هدف پیش‌بینی برای مدل انتشار عمل می‌کند. برای اطمینان از کیفیت این نمایش نهان، از یک مدل ازپیش‌آموزش‌دیده مبتنی بر ImageNet (Russakovsky et al., 2015) استفاده شده است. این فشرده‌سازی به کاهش حجم داده و ساده‌سازی فرآیند یادگیری کمک می‌کند (Rombach et al., 2022).

Visual Attention Conditioning (VAC): ماژول VAC برای هدایت فرآیند حذف نویز در مدل انتشار طراحی شده است. این ماژول از مدل خودتوجهی VIT B-32 ارائه‌شده توسط PyTorch استفاده می‌کند (Dosovitskiy et al., 2021). این ماژول، فریم سیاه و سفید را به یک بردار ویژگی $V_{features} \in R^{50\times 768}$ تبدیل می‌کند که به‌طور مستقیم در لایه‌های کانولوشن مدل انتشار وارد شده و نقش راهنما را برای تولید رنگ ایفا می‌کند.

Diffusion Model Latent Colorization (DMLC): این بخش، هسته اصلی مدل است. این ماژول با استفاده از توزیع نویزی تصادفی و بردار ویژگی VAC، فضای نهان رنگی Latc را تولید می‌کند. برای این کار، از معماری U-Net الهام گرفته‌ایم (Ronneberger et al., 2015). مدل انتشار به‌جای تولید مستقیم پیکسل، فضای نهان را تولید نموده که محاسبات را سبک‌تر و قابل کنترل‌تر می‌کند.

Decoder: رمزگشا[2] برای بازسازی فریم رنگی از فضای نهان Latc استفاده می‌شود. مشابه بخش رمزگذار، این بخش نیز از مدل ازپیش‌آموزش‌دیده بهره می‌برد.

---
1. latent space
2. decoder



شکل ۱. توپولوژی شبکه پیشنهادی ما که جریان داده را در مراحل آموزش و استنتاج نشان می‌دهد. فریم سیاه و سفید S_g به نسخه رنگی‌شده S_c تبدیل می‌شود. این تبدیل از طریق خروجی LatR در ماژول DMLC انجام شده و سپس، توسط رمزگشا D بازسازی می‌شود (در جریان آموزش مسیر).

۴-۳- آموزش

برای تقویت عملکرد اجزاء مدل، تنها ماژول DMLC به‌صورت خاص آموزش داده شد تا تأثیر مستقل مدل انتشار بر فرایند رنگی‌سازی بررسی شود.

در هر مرحله آموزش، از دسته‌های ۱۰۰ فریمی استفاده کردیم و آموزش را درمجموع برای ۳۰۰ دوره (epoch) انجام دادیم. به‌عنوان بهینه‌ساز، از AdamW (Loshchilov & Hutter, 2017) با نرخ یادگیری اولیه e-52 استفاده شد که هر ۵۰ دوره با ضریب ۰/۱ کاهش می‌یافت. تابع خطای مورد استفاده، میانگین مربعات خطا (MSE) بود که در هر دو مجموعه داده به کار گرفته شد. تمام آزمایش‌ها روی سیستمی با ویندوز ۱۱، پردازنده AMD Ryzen 5600g، رم ۳۲ گیگابایتی و کارت گرافیک NVIDIA GTX 1080 Ti انجام شد.

۴-۴-آموزش

در مرحله استنتاج[1]، مدل تنها به یک فریم سیاه و سفید (S_g) به‌عنوان ورودی نیاز دارد و خروجی آن، فریم رنگی‌شده متناظر (S_c) است. ارزیابی نهایی با مقایسه Sc با فریم رنگی مرجع R انجام می‌شود.

1. inference

۵- معیارهای ارزیابی

در این پژوهش، برای ارزیابی کمی کیفیت ویدئوهای رنگی‌شده، فریم‌های تولیدشده را با فریم‌های اصلی رنگی[1] مقایسه کردیم. این مقایسه به ما اجازه می‌دهد تا اختلاف بین خروجی مدل و مرجع واقعی را از منظر کیفیت عددی بررسی کنیم.

برای این منظور، از چندین معیار متداول در حوزه بینایی ماشین استفاده کردیم: نسبت سیگنال به نویز اوج (PSNR)، شاخص شباهت ساختاری (SSIM)، فاصله فریشه در فضای ویژگی‌ها (FID) و شاخص یکنواختی توزیع رنگ (CDC). در ادامه، نحوه محاسبه و تفسیر هر یک از این معیارها توضیح داده می‌شود.

۵-۱- معیارهای ارزیابی

معیار PSNR مستقیماً بر اساس شدت پیکسل‌ها عمل می‌کند و با مقایسه حداکثر مقدار ممکن شدت با خطای میانگین مربعات (MSE) محاسبه می‌شود. فرمول آن به‌صورت زیر تعریف می‌شود:

$$PSNR = 10 \times log_{10} \left(\frac{(Lmax)^2}{MSE} \right) \quad (۱)$$

که در آن، L_{max} حداکثر مقدار ممکن شدت پیکسل‌هاست. هرچه مقدار PSNR بالاتر باشد، شباهت بیشتری میان تصویر تولیدشده و تصویر مرجع وجود دارد.

با اینکه PSNR یک معیار کمی شناخته‌شده برای کیفیت تصویر است؛ اما لزوماً با درک بصری انسان مطابقت کامل ندارد؛ چراکه ساختار و الگوهای تصویری را در نظر نمی‌گیرد (Zhang et al., 2012).

بااین‌حال، سادگی و قابلیت تفسیر بالای PSNR باعث شده که در مقایسه‌های فریم‌به‌فریم در ویدئو، به‌ویژه در پژوهش‌های DLVC، به‌طور گسترده استفاده شود (Jampour et al., 2022; Kouzouglidis et al., 2019; Shi et al., 2020).

۵-۲- شاخص شباهت ساختاری[2]

برخلاف PSNR، معیار SSIM اطلاعات ساختاری، کنتراست و روشنایی را نیز در نظر می‌گیرد و به همین دلیل، شباهت بیشتری به درک انسانی دارد (Wang et al., 2004).

فرمول SSIM به‌صورت زیر تعریف می‌شود:

$$SSIM(a,b) = \frac{\left((2\mu a\mu b + C^1)(2\sigma ab + C^2)\right)}{\left((\mu a^2 + \mu b^2 + C^1)(\sigma a^2 + \sigma b^2 + C^2)\right)} \quad (۲)$$

که در آن:

- μa و μb، میانگین شدت تصاویر a و b هستند.
- σa و σb، انحراف معیار آن‌هاست.
- σab، کوواریانس بین دو تصویر است.
- C_1 و C_2، پارامترهایی برای پایداری محاسبه‌اند که معمولاً بر پایه مقادیر $k_1 = 0.001$ و $k_2 = 0.003$ تعریف می‌شوند.

در این پژوهش، SSIM برای مقایسه کیفیت بصری بین خروجی رنگی و تصویر مرجع استفاده شد.

---
1. ground truth
2. Structural similarity index metric (SSIM)

۵-۳- فاصله فریشه در فضای ویژگی‌ها

FID، یکی از معیارهای محبوب برای ارزیابی کیفیت تصاویر تولیدشده توسط مدل‌های مولد است (Heusel et al., 2017). این معیار به‌جای مقایسه پیکسل‌به‌پیکسل، ویژگی‌های سطح بالای تصویر را در فضای شبکه عصبی (اغلب با استفاده از مدل InceptionV3) مقایسه می‌کند.

مقدار فاصله فریشه کمتر، نشان‌دهنده شباهت بیشتر بین تصاویر تولیدشده و نمونه‌های واقعی است. این ویژگی باعث شده که FID، به‌ویژه در سناریوهای نظارت‌شده یا نیمه‌نظارتی که فریم مرجع در دسترس است، معیار مناسبی برای ارزیابی کیفیت رنگی‌سازی در DLVC محسوب شود.

۵-۴- فاصله فریشه در فضای ویژگی‌ها

اگرچه FID کیفیت تصویر را از منظر ادراک بصری بررسی می‌کند؛ اما برای ارزیابی یکنواختی انتقال رنگ در طول فریم‌های ویدئو به معیار مکملی نیاز داریم. برای این منظور، از شاخص یکنواختی توزیع رنگ (CDC) استفاده شده است.

CDC بر پایه شاخص جِنسن-شانون (JS) عمل می‌کند که فاصله بین توزیع هیستوگرام رنگ فریم‌های متوالی را محاسبه می‌کند. فرمول کلی به‌صورت زیر است:

$$CDC_{\mathrm{t}} = \frac{1}{3 \times (N - t)} \sum_{c \in \{r,g,b\}} \sum_{i=1}^{N-t} JS\left(P_c(I^i), P_c(I^{i+t})\right) \quad (۳)$$

که در آن:

- N، تعداد کل فریم‌ها در ویدئو است.
- $P_c(I^i)$، توزیع احتمال نرمال‌شده رنگ (r, g, b) برای تصویر i است.
- t، فاصله زمانی بین فریم‌هاست (برای مثال t = 1، ۲ یا ۳).

برای ارزیابی جامع، مقادیر CDC برای tهای مختلف میانگین‌گیری شده‌اند:

$$CDC = \frac{(CDC^1 + CDC^2 + CDC^4)}{3} \quad (۴)$$

مقدار پایین‌تر CDC نشان‌دهنده یکنواختی بهتر در انتشار رنگ بین فریم‌ها در طول زمان است.

۶- نتایج تجربی

پس از پایان مرحله آموزش، عملکرد مدل ما روی مجموعه‌ای از ویدئوهایی که در داده‌های آموزشی حضور نداشتند، ارزیابی شد. این مجموعه شامل بخش آزمون DAVIS، بخش اعتبارسنجی مجموعه LDV و داده‌های مورد استفاده در چالش NTIRE بود. در ادامه، نتایج به‌دست‌آمده از نظر عددی (ارزیابی کمی) و همچنین بصری (ارزیابی کیفی) ارائه و تحلیل می‌شوند.

۶-۱- ارزیابی کمی

برای ارزیابی عملکرد عددی مدل، از چهار معیار PSNR، SSIM، FID و CDC استفاده کردیم. جدول ۱، نتایج مدل ما را با سایر روش‌های موجود بر روی مجموعه داده DAVIS مقایسه می‌کند.

جدول ۱. مقایسه نتایج مدل پیشنهادی با روش‌های موجود روی مجموعه داده DAVIS

روش	PSNR	SSIM	CDC	FID
لی و همکاران (۲۰۱۹)	30.35	–	–	6.22e−4
چن و همکاران (۲۰۲۳)	–	–	4.02e−3	5.87e−4
هوانگ و همکاران (۲۰۲۲)	30.61	–	–	6.87e−4
مدل پیشنهادی ما	27.95	0.27	3.19e−3	5.02e−4

شکل ۲. نتایج استنتاج مدل ما بر روی مجموعه داده DAVIS که کیفیت رنگی‌سازی فریم‌ها را نشان می‌دهد. درحالی‌که در برخی موارد نشت رنگ مشاهده می‌شود، نتایج نهایی، شباهت زیادی به تصاویر رنگی اصلی دارند.

با وجود اینکه مقدار PSNR و SSIM مدل ما نسبت به برخی روش‌ها کمتر است؛ اما کاهش قابل‌توجه مقدار FID و CDC نشان‌دهنده بهبود کیفیت درک‌شده تصویر و ثبات رنگی در طول فریم‌هاست. این نتایج، اثربخشی مدل پیشنهادی ما در حفظ یکپارچگی بصری و رنگی فریم‌های ویدئویی را تأیید می‌کنند.

۶-۲- فاصله فریشه در فضای ویژگی‌ها

ارزیابی کیفی از طریق بررسی بصری فریم‌های رنگی‌شده صورت گرفت و نتایج با فریم‌های رنگی اصلی مقایسه شد. در شکل ۲، نتایج مدل ما روی مجموعه داده DAVIS نمایش داده شده است. در این تصویرها، سه حالت برای هر فریم ارائه شده‌اند:

۱. فریم رنگی مرجع (R)

۲. نسخه سیاه و سفید آن (Sg)

۳. خروجی تولیدشده توسط مدل ما (Sc)

شکل ۲ نشان می‌دهد که مدل قادر به تولید رنگ‌هایی نزدیک به نسخه واقعی بوده و در اکثر موارد، یکنواختی رنگ را در طول تصویر حفظ کرده است. اگرچه در برخی نمونه‌ها نشت رنگ[1] دیده می‌شود؛ اما درمجموع، خروجی مدل از نظر بصری، رضایت‌بخش و واقع‌گرایانه است.

۷- فاصله فریشه در فضای ویژگی‌ها

نتایج این پژوهش نشان می‌دهد که استفاده از مدل‌های مبتنی بر احتمالات و انتشار برای رنگی‌سازی ویدئو، رویکردی مؤثر و کارآمد است که می‌تواند کیفیت رنگ را در فریم‌های ویدئویی به‌خوبی بازسازی کند. مدل ارائه‌شده در مقایسه با تصاویر مرجع، دقت رنگی مناسبی از خود نشان داده و برای کاربردهایی مانند بازسازی فیلم‌های قدیمی یا بازآفرینی محتوای تاریخی، گزینه‌ای مناسب به نظر می‌رسد.

معماری پیشنهادی و ارزیابی‌های جامع، گامی روبه‌جلو در بهره‌گیری از مدل‌های انتشار در مسائل DLVC به شمار می‌آید. امید است یافته‌های این تحقیق منجر به توسعه مدل‌های بهینه‌تر و قدرتمندتری در آینده نزدیک شود.

درنهایت، پژوهش ما ثابت می‌کند که تکنیک انتشار شرطی می‌تواند راهکاری نویدبخش برای رنگی‌سازی خودکار ویدئوهای سیاه و سفید باشد. شایان ذکر است که تمام نتایج گزارش‌شده با منابع سخت‌افزاری محدود حاصل شده‌اند که قابلیت پیاده‌سازی عملی مدل را تقویت می‌کند.

منابع

Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18208-18218).

Ceylan, D., Huang, C. H. P., & Mitra, N. J. (2023). Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 23206-23217).

Chen, S., Li, X., Zhang, X., Wang, M., Zhang, Y., Han, J., & Zhang, Y. (2024). Exemplar-based video colorization with long-term spatiotemporal dependency. *Knowledge-Based Systems*, *284*, 111240.

Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10850-10869.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, *34*, 8780-8794.

Dogan, P., Aydın, T. O., Stefanoski, N., & Smolic, A. (2015). Key-frame based spatiotemporal scribble propagation. In *Eurographics Workshop on Intelligent Cinematography and Editing* (pp. 13–20).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

1. color leaking

Furusawa, C., Kitaoka, S., Li, M., & Odagiri, Y. (2021). Generative probabilistic image colorization. *arXiv preprint arXiv:2109.14518*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Heu, J. H., Hyun, D. Y., Kim, C.-H., & Lee, S. U. (2009). Image and video colorization based on prioritized source propagation. In *16th IEEE International Conference on Image Processing* (pp. 465–468).

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, *30*.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., et al. (2022). Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840-6851.

Huang, R., Li, S., Dai, W., Li, C., Zou, J., & Xiong, H. (2022). Improving optical flow inference for video colorization. In *IEEE International Symposium on Circuits and Systems* (pp. 3185–3189).

Iizuka, S., & Simo-Serra, E. (2019). DeepRemaster: Temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics*, *38*(6), 1-13.

Jampour, M., Zare, M., & Javidi, M. (2022). Advanced Multi-GANs towards near to real image and video colorization. *Journal of Ambient Intelligence and Humanized Computing*, 1-18.

Kingma, D.P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kouzouglidis, P., Sfikas, G., & Nikou, C. (2019). Automatic video colorization using 3D conditional generative adversarial networks. In *International Symposium on Visual Computing* (pp. 209–218). Springer.

Lei, C., & Chen, O. (2019). Fully automatic video colorization with self-regularization and diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3753–3761).

Liu, Y., Zhao, H., Chan, K.C.K., Wang, X., Loy, C.C., Qiao, Y., & Dong, C. (2021). Temporally consistent video colorization with deep feature propagation and self-regularization learning. *arXiv preprint arXiv:2110.04562*.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mahajan, A., Patel, N., Kotak, A., & Palkar, B. (2021). An end-to-end approach for automatic and consistent colorization of gray-scale videos using deep-learning techniques. In *International Conference on Machine Intelligence and Data Science Applications* (pp. 539–551). Springer.

Meyer, S., Cornillere, V., Djelouah, A., Schroers, C., & Gross, M. (2018). Deep video color propagation. *arXiv preprint arXiv:1808.03232*.

Preechakul, K., Chatthee, N., Wizadwongsa, S., & Suwajanakorn, S. (2022). Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211-252.

Salmona, A., Bouza, L., & Delon, J. (2022). DeOldify: A review and implementation of an automatic colorization method. *Image Processing OnLine*, 12, 347–368.

Shi, M., Zhang, J. Q., Chen, S. Y., Gao, L., Lai, Y. K., & Zhang, F.L. (2020). Deep line art video colorization with a few references. *arXiv preprint arXiv:2003.10685*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning* (pp. 2256–2265). PMLR.

Stival, L., & Pedrini, H. (2023). Survey on video colorization: Concepts, methods and applications. *Journal of Signal Processing Systems*, 1-24.

Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Leibe, B., & Chen, L. C. (2019). FEELVOS: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9481–9490).

Wan, Z., Zhang, B., Chen, D., Zhang, P., Chen, D., Liao, J., & Wen, F. (2020). Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2747–2757).

Wang, Z., Lu, L., & Bovik, A. C. (2004). Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, *19*(2), 121-132.

Xu, X., Wang, Z., Zhang, G., Wang, K., & Shi, H. (2023). Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 7754–7765).

Yang, R., & Timofte, R. (2021). NTIRE 2021 Challenge on quality enhancement of compressed video: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Yang, R., Srivastava, P., & Mandt, S. (2022). Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*.

Yatziv, L., & Sapiro, G. (2006). Fast image and video colorization using chrominance blending. *IEEE Transactions on Image Processing*, *15*(5), 1120-1129.

Zhang, B., He, M., Liao, J., Sander, P.V., Yuan, L., Bermak, A., & Chen, D. (2019). Deep exemplar-based video colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8052–8061).

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2012). A comprehensive evaluation of full reference image quality assessment algorithms. In *19th IEEE International Conference on Image Processing* (pp. 1477–1480). IEEE.

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., & Feng, J. (2022). MagicVideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.

استناد به این مقاله: حضرتی فرد، سید مهدی، و چکینی، امیر. (۱۴۰۴). رنگی‌سازی ویدئو بر پایه پیاده‌سازی مدل انتشار. *فصلنامه پژوهش‌های نوین در شهر هوشمند*، ۳(۳)، ۶-۱۸.