

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

گزارش پروژه پایانی درس بینایی ماشین

آشنایی با مدل‌های پخشی (Diffusion)

نگارش

غلامرضا دار

استاد درس

دکتر رضا صفابخش

زمستان ۱۴۰۱

چکیده

مدل‌های پخشی یک دسته جدید از مدل‌های مولد هستند که در طی سال‌های اخیر به یک رقیب جدی برای مدل‌های مولد تخاصمی یا خودکدگزارهای تغییراتی شده‌اند. این مدل‌ها در دو مرحله مستقیم و عکس عمل می‌کنند به این شکل که ابتدا به تصویر ورودی نویز اعمال می‌کنند و در مرحله عکس سعی می‌کنند این نویز را کاهش بدهند و به تصویر اصلی برسند. در این پروژه تحقیقاتی به بررسی تکامل این دسته مدل‌ها در طی چند سال اخیر و همچنین کاربردهایی که می‌توانند داشته باشند می‌پردازیم.

کلمات کلیدی:

شبکه‌های عصبی عمیق، مدل‌های مولد، مدل‌های پخشی

فهرست مطالب

۱	۱	۱ مقدمه
۲	۲	۲ مدل های پخشی (Diffusion)
۳	۲-۱	۲-۱ نحوه عملکرد مدل های پخشی
۵	۲-۲	۲-۲ مجموعه داده و معیارهای ارزیابی
۷	۳-۲	۳-۲ نتایج و خواص مدل های پخشی
۷	۳-۲-۱	۳-۲-۱ نتایج
۹	۳-۲-۲	۳-۲-۲ خواص
۱۱	۳	۳ بهبود مدل های پخشی
۱۲	۳-۱	۳-۱ بهبود مدل در معیار درست نمایی
۱۴	۳-۲	۳-۲ کاهش تعداد تکرارها
۱۵	۳-۳	۳-۳ مقیاس پذیری مدل
۱۶	۴	۴ کاربردهای مختلف مدل های پخشی
۱۷	۴-۱	۴-۱ Palette
۱۸	۴-۱-۱	۴-۱-۱ رنگی سازی تصاویر خاکستری
۱۸	۴-۱-۲	۴-۱-۲ بازسازی تصویر
۱۹	۴-۱-۳	۴-۱-۳ گسترش تصویر
۲۰	۴-۱-۴	۴-۱-۴ رفع آسیب های JPEG
۲۱	۴-۱-۵	۴-۱-۵ نتیجه گیری
۲۲	۴-۲	۴-۲ DALL-E
۲۴	۴-۲-۱	۴-۲-۱ جزیيات روش تولید تصاویر
۲۴	۴-۲-۲	۴-۲-۲ نتایج و ویژگی ها
۲۶	۴-۲-۳	۴-۲-۳ اهمیت مدل پیشین
۲۶	۴-۲-۴	۴-۲-۴ نتیجه گیری

۲۷	Stable Diffusion ۳-۴
۲۷	۱-۳-۴ معماری و روش
۲۹	۲-۳-۴ نتایج و ویژگی‌ها
۳۱	Dreambooth ۴-۴
۳۸	۵ بحث و جمع بندي
۳۹	فهرست مراجع

فهرست اشکال

..... ۱ شکل ۱ - شمایی از ساختار کلی تعدادی از مدلهای مولد رایج
..... ۲ شکل ۲ - تعدادی از تصاویر تولید شده توسط مدلهای پخشی [۱]
..... ۳ شکل ۳ - ساختار کلی مراحل مستقیم و عکس مدلهای پخشی
..... ۵ شکل ۴ - تعدادی تصویر از بخش اتاق خواب مجموعه داده LSUN
..... ۶ شکل ۵ - نیازمندیهای لازم برای معیار IS
..... ۸ شکل ۶ - تصاویر تولید شده توسط مدل پخشی برای مجموعه داده LSUN-church
..... ۸ شکل ۷ - تصاویر تولید شده توسط مدل پخشی برای مجموعه داده LSUN-bedroom
..... ۹ شکل ۸ - درونیابی نهان به کمک مدل های پخشی
..... ۱۰ شکل ۹ - استفاده از تصاویر میانی برای ایجاد تصاویر مشابه
..... ۱۲ شکل ۱۰ - تغییر نسبت واریانس آموخته شده به واریانس ثابت در طی تکرارهای مختلف
..... ۱۳ شکل ۱۱ - جملات تابع زیان نسبت به تکرار
..... ۱۳ شکل ۱۲ - زمانبند خطی(بالا) و زمانبند کسینوسی(پایین)
..... ۱۴ شکل ۱۳ - نمونه برداری با تکرارهای مختلف
..... ۱۵ شکل ۱۴ - بررسی مقیاس پذیری مدل پخشی
..... ۱۶ شکل ۱۵ - تعدادی از مثال های تبدیل متن به تصویر به کمک مدل 2 DALL-E
..... ۱۷ شکل ۱۶ - کاربردهای مختلف روش پالت
..... ۱۸ شکل ۱۷ - رنگی کردن تصاویر خاکستری
..... ۱۹ شکل ۱۸ - مقایسه روش های مختلف در مسئله ترمیم تصویر
..... ۱۹ شکل ۱۹ - نتیجه گسترش تصویر به کمک روش پالت
..... ۲۰ شکل ۲۰ - مقایسه نتایج گسترش تصویر به کمک روش پالت و سایر روش ها
..... ۲۱ شکل ۲۱ - نتیجه کاهش آسیب های JPEG
..... ۲۲ شکل ۲۲ - نمونه تصویر تولید شده توسط دال ای ۲ به همراه متن ورودی
..... ۲۳ شکل ۲۳ - شمای کلی معماری دال ای ۲
..... ۲۵ شکل ۲۴ - تولید تصاویر مشابه از نظر معنایی به کمک بردar معنایی CLIP
..... ۲۵ شکل ۲۵ - تعدادی از تصاویر تولید شده توسط دال ای ۲

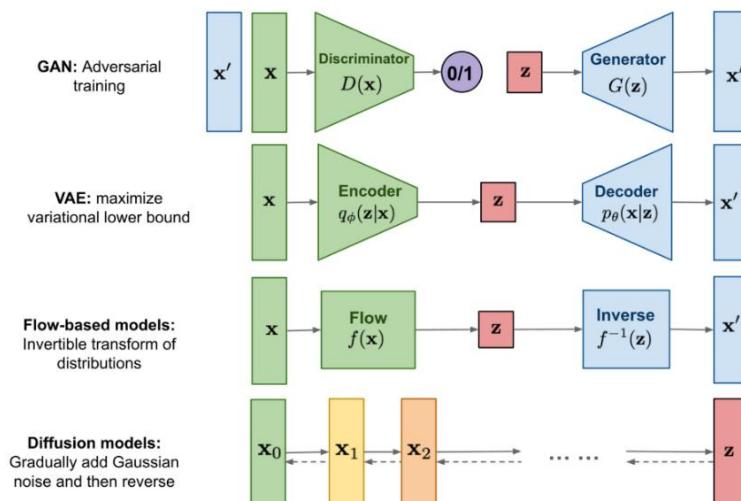
۲۷.....	شکل ۲۶-معماری کلی روش LDM
۲۸.....	شکل ۲۷- مقایسه اثر میزان کاهش ابعاد بر معیار FID (کمتر=بهتر)
۲۸.....	شکل ۲۸ - مقایسه اثر میزان کاهش ابعاد بر معیار IS بر حسب میزان آموزش (بیشتر=بهتر)
۲۹.....	شکل ۲۹ - برخی از تصاویر تولید شده توسط روش LDM
۲۹.....	شکل ۳۰ - برخی از تصاویر تولید شده با ورودی متن توسط روش LDM
۳۰.....	شکل ۳۱ - نتیجه ترمیم تصویر به کمک روش LDM
۳۰.....	شکل ۳۲ - نتیجه تولید تصویر بر اساس برچسب معنایی پیکسل ها
۳۱.....	شکل ۳۳ - نتیجه تولید تصویر به کمک برچسب کلاس نواحی ای از تصویر
۳۲.....	شکل ۳۴ - تصاویر تولید شده شخصی سازی شده با کمک روش DreamBooth
۳۳.....	شکل ۳۵ - شمای کلی سیستم DreamBooth
۳۴.....	شکل ۳۶ - نحوه آموزش مدل DreamBooth
۳۵.....	شکل ۳۷ - مقایسه روش های مختلف تولید تصویر با سوژه مشخص
۳۶.....	شکل ۳۸ - قابلیت تغییر ویژگیهای یک تصویر در مدل DreamBooth
۳۶.....	شکل ۳۹ - قابلیت تغییر ویژگیهای یک تصویر در مدل DreamBooth
۳۶.....	شکل ۴۰ - تغییر نمای تصویر و ایجاد تصاویری از نماهای بدیع
۳۷.....	شکل ۴۱ - برخی مشکلات و محدودیتهای مدل DreamBooth

فهرست جداول

جدول ۱ - مقایسه مدل های مختلف تولید تصویر.....	۷
جدول ۲ - مقایسه روش پالت با سایر روش ها در زمینه رنگی کردن تصویر.....	۱۸
جدول ۳ - مقایسه عملکرد روش دال ای ۲ و سایر روش ها.....	۲۵
جدول ۴ - مقایسه روش LDM و سایر روش ها در تولید تصویر بدون شرط.....	۳۱
جدول ۵ - مقایسه روش LDM و سایر روش ها در تولید تصویر با ورودی متن.....	۳۱

۱ مقدمه

در سال‌های اخیر، مدل‌های مولد عمیق^۱ مانند مدل‌های مولد تخصصی یا GAN^۲، خودکدگذارهای تغییراتی یا VAE^۳، مدل‌های Autoregressive و مدل‌های بر اساس جریان^۴ توانسته اند در زمینه تولید تصویر^۵ و صوت به نتایج خیره‌کننده ای دست یابند. این مدل‌های مولد به طور کلی یک مدل آماری هستند که وظیفه تبدیل نویز به داده از یک توزیع مورد نظر را بر عهده دارد. یکی از دست‌آوردهای دیگر که اخیراً توجه محققان و عموم را به خود جلب کرده است، مدل‌های پخشی^۶ نام دارند. انتشار مدل‌هایی مانند DALL-E 2، GLIDE، Imagen و Stable Diffusion قدرت مدل‌های پخشی در تولید تصاویر را به خوبی نشان داده‌اند و آن‌ها را به یک رقیب جدی برای GAN‌ها و VAE‌ها تبدیل کرده‌اند.



شکل ۱ - شمایی از ساختار کلی تعدادی از مدل‌های مولد رایج

در این پژوهه ابتدا با این دسته از مدل‌ها آشنا می‌شویم، نقاط قوت و ضعف آن‌ها را بررسی می‌کنیم، راههایی برای بهبود این ضعف‌ها ارائه می‌دهیم، کاربردهای مختلف این مدل‌ها را در زمینه‌های مختلف بررسی می‌کنیم و در نهایت با جمع‌بندی در مورد کل پژوهه، آینده زمینه تولید تصویر و مدل‌های پخشی را بررسی می‌کنیم.

¹ Deep Generative Models

² Generative Adversarial Network

³ Variational Autoencoder

⁴ Flow

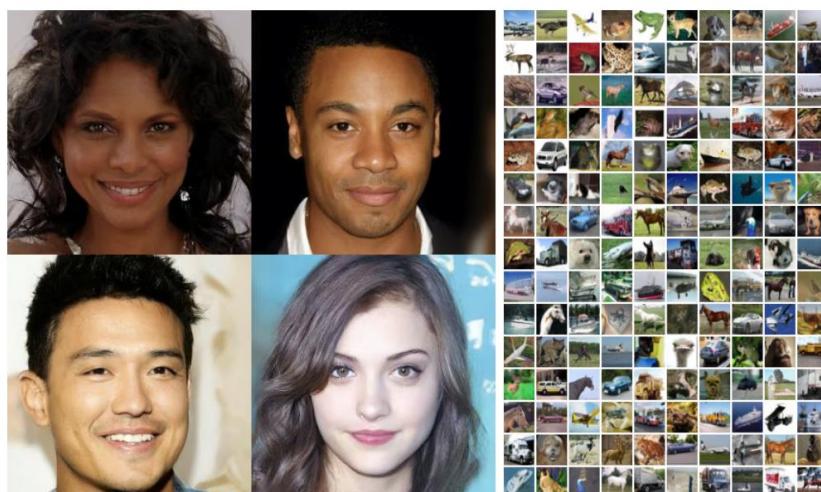
⁵ Image Generation

⁶ Diffusion Models

۲ مدل های پخشی (Diffusion)

همان طور که در فصل مقدمه نیز گفته شد، در سال های اخیر مدل های مولد عمیق، در کاربرد تولید تصویر، نتایج بسیار باور نکردنی ای داشته اند. مدل های رایج مخصوصاً مدل های بر پایه مولد تخاصمی، علی رغم عملکرد بسیار عالی در زمینه تولید تصویر، مشکلاتی اعم از فروپاشی میانه^۷ را دارند. این مشکل به این صورت است که زمانی که بخش مولد^۸ مدل، قادر می شود تصاویری تولید کند که بخش متمایزگر^۹ مدل را فربیب دهد، متوقف می شود و انگیزه ای برای ادامه یادگیری نخواهد داشت و به این ترتیب تنوع تصاویر تولید شده محدود می گردد.

در سال ۲۰۲۰، Ho و همکاران در مقاله [۱] Denoising Diffusion Probabilistic Models این مدل جدید به اسم مدل پخشی، به عنوان رقیب برای مدل های مولد تخاصمی و خودکدگزار تغییراتی در زمینه تولید تصویر معرفی کردند. اخیراً محققان علاقه زیادی به این مدل های پخشی نشان داده اند و عملکرد آن ها را در امور مختلف مورد آزمایش قرار داده اند. در این فصل به طور دقیق تر با این مدل ها آشنا می شویم



شکل ۲ – تعدادی از تصاویر تولید شده توسط مدل های پخشی [۱]

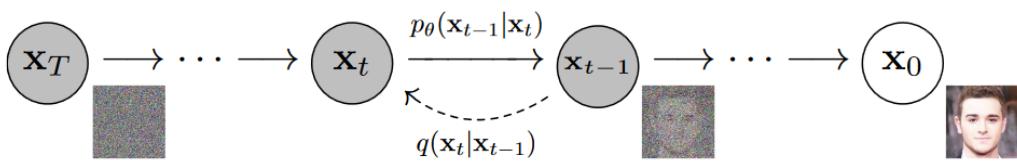
⁷ Mode Collapse

⁸ Generator

⁹ Discriminator

۲-۱ نحوه عملکرد مدل‌های پخشی

به طور کلی مدل‌های پخشی شامل دو مرحله مستقیم^{۱۰} و عکس^{۱۱} هستند. ایده اصلی این مدل‌ها این است که به تصویر ورودی، در مرحله مستقیم، طی چندین تکرار^{۱۲}، نویز افروده می‌شود سپس در مرحله عکس، سعی می‌شود تابعی تخمین زده شود که بتواند این تصویر نویزی را به تصویر اصلی بدون نویز بازگردداند.



شکل ۳- ساختار کلی مراحل مستقیم و عکس مدل‌های پخشی

اگر به شکل ۳ توجه کنید، دو توزیع p و q را مشاهده می‌کنید. توزیع q در واقع توزیع مربوط به مرحله مستقیم است. توسط این توزیع، تصویر x_t نویزی شده و به تصویر x_{t+1} تبدیل می‌شود. این پروسه نویزی شدن تصویر به تعداد T دفعه تکرار می‌شود (معمولا $T = 1000$) به شکلی که تصویر x_T کاملا نویز گاوی خواهد بود. توزیع p_θ مربوط به مرحله عکس یا کاهش نویز مدل است. این توزیع تصویر x_t را می‌گیرد و تصویر کم‌نویزتر x_{t-1} را تخمین می‌زند.

اگر به نحوه نوشتن p_θ, q دقت کنید، متوجه می‌شوید که توزیع p_θ به پارامتر θ وابسته است و در واقع یک شبکه عصبی قابل آموزش است اما توزیع q پارامتری ندارد و در مرحله مستقیم توسط یک زمانبند نویز^{۱۳} از پیش تعیین شده، در هر مرحله مقداری نویز به تصویر مرحله قبل افزوده می‌شود. نویسنده‌های این مقاله با انجام یک بازپارامترسازی^{۱۴}، مرحله عکس را به نوعی تغییر داده اند که هدف این مرحله گرفتن یک تصویر و تخمین زدن مقدار نویز موجود در آن باشد و نه بازسازی تصویر اصلی که طبق گفته نویسنده‌ها آموزش این مدل را آسان تر می‌کند.

به طور کلی برای تولید تصاویر به کمک مدل‌های پخشی دو روش وجود دارد، تولید تصاویر به صورت شرطی و بدون شرط. تولید تصاویر بدون شرط به این معناست که مدل صرفا نویز ورودی را به داده نمایانگر یک توزیع موردنظر تبدیل می‌کند. پروسه تولید تصویر در این روش، توسط عوامل خارجی کنترل نمی‌شود و مدل آزادانه می‌تواند هر تصویری را تولید کند. تولید تصویر به صورت شرطی به این شکل عمل می‌کند که علاوه بر نویز ورودی، به مدل یک یا چند ورودی دیگر نیز داده می‌شود. این ورودی می‌تواند نسخه کدشده یک متن یا حتی یک تصویر دیگر باشد و با دادن این

¹⁰ forward

¹¹ reverse

¹² iteration

¹³ Noise scheduler

¹⁴ Reparametrization

ورودی‌ها به مدل، انتظار می‌رود تصاویر تولید شده توسط مدل، به یک زیرمجموعه مشخص محدود شوند که این زیرمجموعه تنظیم ورودی‌های اضافه کنترل می‌شود. به عنوان مثال اگر نسخه کد شده متن "سیب" را به عنوان ورودی اضافه به مدل دهیم، مدل سعی می‌کند تصاویری که تولید می‌کند از نوع سیب باشند. در فصل‌های آینده بیشتر با این گونه تولید تصاویر آشنا می‌شویم.

۲-۲ مجموعه داده و معیارهای ارزیابی

پیش از آنکه بتوانیم نتایج مربوط به مدل‌های پخشی را در زمینه تولید تصویر، با سایر روش‌ها مقایسه کنیم، لازم است با نحوه انجام این مقایسه‌ها و مجموعه داده‌های مورد نیاز آشنا شویم.



شکل ۴ - تعدادی تصویر از بخش اتاق خواب مجموعه داده LSUN

در مقاله ذکر شده، سه مجموعه داده CIFAR10، LSUN و CelebAHQ مورد استفاده قرار گرفته‌اند. مجموعه داده CIFAR10 یک مجموعه داده شامل ۶۰ هزار تصویر ۳۲ پیکسل در ده دسته مختلف است. مجموعه داده LSUN شامل تعدادی زیادی تصویر در دسته بندی‌های مختلف مانند اتاق خواب، کلیسا و تصاویر حیوانات است. مجموعه داده CelebAHQ شامل ۳۰ هزار تصویر ۱۰۲۴ در ۱۰۲۴ پیکسل از چهره افراد مشهور است.

یکی از سوالات اصلی که هنگام بررسی عملکرد مدل‌های تولیدکننده تصویر پیش می‌آید این است که عملکرد این مدل‌ها چگونه بررسی می‌شوند؟ واضح است که انتخاب یک معیار قابل محاسبه توسط کامپیوتر برای توسعه این مدل‌ها ضروری است اما از جهتی دیگر اگر این معیارها با دقت کافی انتخاب نشوند ممکن است مدل‌هایی تولید شوند که از نظر معیار ذکر شده بی‌نظیر باشند اما تصاویر تولیدی آن‌ها از نظر انسان بسیار نامطلوب باشد. بنابراین هنگام تعیین یک معیار ارزیابی برای این مسئله، باید نزدیکبودن به نظر انسان را نیز مورد توجه قرار داد.

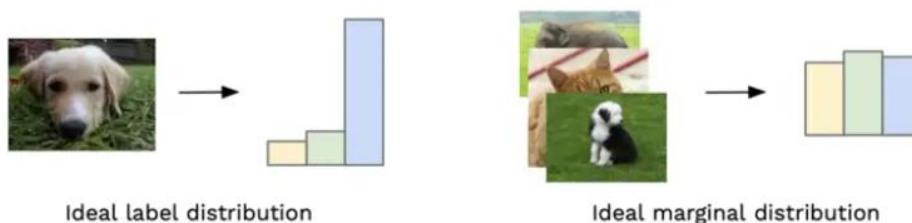
دو معیار مطرحی که در بسیاری از مقالات این زمینه گزارش می‌شود معیارهای IS¹⁵ و FID¹⁶ هستند. معیار IS همان‌طور که از نام آن نیز پیداست، از مدل Inception که یک مدل دسته‌بندی تصاویر است استفاده می‌کند. تصاویر تولید شده توسط یک مدل مولد، به این مدل Inception داده می‌شوند و تعیین می‌شود که تصویر تولید شده با چه احتمالی به هر دسته تصویر تعلق دارد. ایده اصلی روش IS این است که یک مدل مولد خوب، مدلی است که تصاویر تولیدشده

¹⁵ Inception Score

¹⁶ Frechet Inception Distance

توسط آن، توسط مدل Inception به احتمال بالایی فقط به یک دسته تعلق داشته باشند. به عنوان مثال اگر یک تصویر از یک سگ تولید شده است، انتظار می‌رود پیش‌بینی مدل Inception برای آن تصویر با احتمال بالایی دسته سگ باشد. یکی دیگر از ویژگی‌های مطلوبی که در معیار IS به دنبال آن هستیم، این است که تصاویر تولیدشده تنوع بالایی داشته باشند و به عنوان مثال همگی مربوط به دسته سگ نباشند.

در شکل ۵ می‌توان به خوبی دو نیازمندی لازم برای معیار IS را مشاهده کرد.



شکل ۵ - نیازمندی‌های لازم برای معیار IS

معیار دومی که بررسی خواهیم کرد معیار FID یا Frechet Inception Distance نام دارد. همان‌طور که از نام این معیار پیداست، این معیار نوعی فاصله است. در این معیار به جای محاسبه فاصله در سطح پیکسلی و مقایسه مقادیر پیکسل‌های دو تصویر، از خروجی لایه‌های مخفی مدل Inception استفاده می‌شود. این لایه‌ها، اطلاعات معنایی بیشتری از پیکسل‌ها را در خود دارند و در عین حال از برچسب کلاس مربوطه نیز اطلاعات خامتری ارائه می‌دهند که این‌ها باعث شده این معیار در مقایسه تصاویر شباهت زیادی به سیستم بینایی انسان داشته باشد.

حال که با معیارهای لازم برای مقایسه روش‌های تولید تصویر آشنا شدیم، در بخش بعد به تحلیل و مقایسه روش‌های مختلف تولید تصویر می‌پردازیم و عملکرد مدل‌های پخشی را بررسی می‌کنیم.

۲-۳ نتایج و خواص مدل های پخشی

۲-۳-۱ نتایج

در جدول ۱، می‌توان نتایج مربوط به معیارهای IS و FID را برای مدل‌های مختلف تولید تصویر مشاهده کرد. همان‌طور که واضح است، مدل پخشی ارائه شده در این مقاله در هر دو معیار، یکی از بهترین نتایج را کسب کرده است.

جدول ۱ - مقایسه مدل‌های مختلف تولید تصویر

Model	IS	FID	NLL Test (Train)
Conditional			
EBM [4]	8.30	37.9	
JEM [4]	8.76	38.4	
BigGAN [3]	9.22	14.73	
StyleGAN2 + ADA (v1) [29]	10.06	2.67	
Unconditional			
Diffusion (original) [53]			≤ 5.40
Gated PixelCNN [59]	4.60	65.93	3.03 (2.90)
Sparse Transformer [7]			2.80
PixelIQN [43]	5.29	49.46	
EBM [4]	6.78	38.2	
NCSNv2 [56]			31.75
NCSN [53]	8.87 ± 0.12	25.32	
SNGAN [59]	8.22 ± 0.05	21.7	
SNGAN-DDLS [4]	9.09 ± 0.10	15.42	
StyleGAN2 + ADA (v1) [29]	9.74 ± 0.05	3.26	
Ours (L , fixed isotropic Σ)	7.67 ± 0.13	13.51	≤ 3.70 (3.69)
Ours (L_{simple})	9.46 ± 0.11	3.17	≤ 3.75 (3.72)

در تکمیل جدول ۱، اگر به شکل‌های ۶ و ۷ توجه کنید می‌توان تعدادی از نمونه‌های تولید شده توسط این مدل را مشاهده کنید. نکته جالب در مورد شکل ۶ این است که به دلیل وجود برچسب^{۱۷} بر روی برخی تصاویر آموزش، مدل بر روی تصاویر تولید شده نیز اشکالی شبیه به این برچسب‌ها تولید کرده است.

¹⁷ Watermark



شکل ۶ - تصاویر تولید شده توسط مدل پخشی برای مجموعه داده LSUN-church



شکل ۷ - تصاویر تولید شده توسط مدل پخشی برای مجموعه داده LSUN-bedroom

۲-۳-۲ خواص

مدل‌های پخشی علاوه بر نتایج خوبی که با توجه به معیارهای مختلف به دست آوردن، تعدادی خواص و ویژگی دیگر نیز دارند که این مدل‌ها را برای محققان جذاب کرده است. یکی از این ویژگی‌ها امکان درونیابی^{۱۸} در فضای نهان است.



شکل ۸- درونیابی نهان به کمک مدل‌های پخشی

ترکیب کردن تصاویر در فضای پیکسلی مشکلات زیادی دارد. با توجه به بخش سمت چپ شکل ۸، ترکیب کردن دو تصویر در این فضا به دلیل منطبق نبودن بخش‌های مختلف تصاویر (در مورد چهره می‌توان منطبق نبودن اجزای صورت مانند چشم و دهان را مثال زد)، باعث ایجاد مشکلات متعددی اعم از مات‌شدن یا تولید سایه می‌شود.

به کمک مدل‌های پخشی می‌توان عمل درونیابی را در فضای نهان به شکل زیر انجام داد.

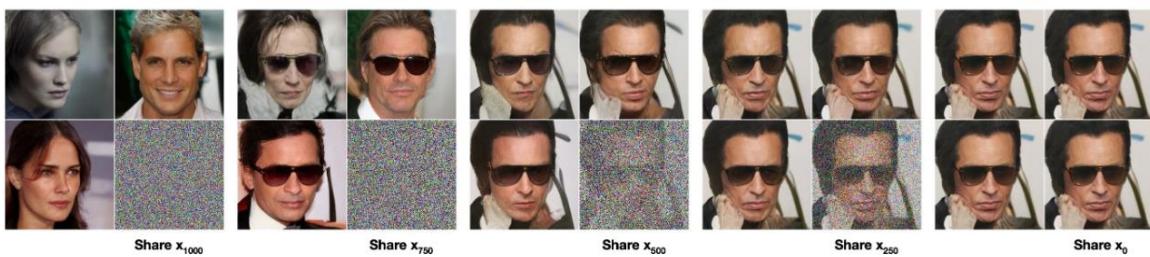
$$\bar{x}_t = (1 - \lambda)x_0 + \lambda x'_0 \quad (1)$$

$$\bar{x}_0 \sim p(x_0 | \bar{x}_t) \quad (2)$$

طبق روابط ۱ و ۲، عمل درونیابی در فضای نهان انجام می‌شود سپس به هنگام نمونه‌برداری، تصویر \bar{x}_0 از توزیعی که توسط بردار نهان درونیابی شده تعیین شده برداشته می‌شود. شکل ۸ نتیجه انجام دادن این کار را نشان می‌دهد. در هر سطر، تصاویر سمت چپ و راست تصاویر اصلی از مجموعه داده CelebAHQ هستند و تصاویر بین، تصاویری هستند که توسط روش توضیح داده شده تولید شده اند. مقدار λ میزان ترکیب شدن دو تصویر چپ و راست را مشخص می‌کند.

¹⁸ Interpolation

یکی دیگر از کاربردهای جالب مدل‌های پخشی، ایجاد شرط بر روی تولید تصاویر به کمک تصاویر میانی (تصاویری که کاملاً نویز نیستند و مقداری کاهش نویز دیده اند) است. در حالت اصلی، تصاویر بر اساس x_T که کاملاً نویز گاووسی است تولید می‌شوند، اما می‌توان یک تصویر میانی مانند x_{750} را به عنوان مبدأ انتخاب کرد و تعدادی تصویر جدید را به کمک این تصویر تولید کرد. نتیجه این است که تصاویر تولید شده جدید، خواص و ویژگی‌های معنایی زیادی را با یکدیگر و با تصویر میانی x_{750} به اشتراک دارند. شکل ۹ این پدیده را به خوبی نشان می‌دهد. تصاویر تولید شده توسط تصویر میانی x_{750} شباهت‌های معنایی زیادی به یکدیگر دارند اما چهره افراد تولید شده با یکدیگر متفاوت است در صورتی که تصاویر تولید شده توسط $x_{T=1000}$ که صرفاً نویز گاووسی است، کاملاً با یکدیگر متفاوت اند.



شکل ۹ - استفاده از تصاویر میانی برای ایجاد تصاویر مشابه

۳ بهبود مدل‌های پخشی

پس از موفقیت‌های مقاله DDPM^[۱]، حجم عظیمی از محققان سعی در بهبود و استفاده از آن داشتند. در این فصل یکی از مقاله‌های تاثیرگذار و مهم که تعداد زیادی از مشکلات مدل‌های پخشی را رفع می‌کند را بررسی خواهیم کرد.

مقاله Improved Denoising Probabilistic Models^[۲] به طور کلی ۳ بهبود اصلی بر روی مقاله اصلی مدل‌های پخشی ارائه کرد. اولین کاری که در این مقاله انجام شد این بود که نویسنده‌های این مقاله با انجام تعدادی تغییر کوچک توانستند یکی از ضعف‌های اصلی مدل‌های پخشی که امتیاز لگاریتم درست‌نمایی^[۱۹] بود را رفع کنند. تا پیش از این یکی از ضعف‌های اصلی مدل‌های پخشی نسبت به مدل‌های مولد خصمانه، معیار لگاریتم درست‌نمایی بود. نویسنده‌های مقاله DDPM به این نتیجه رسیده بودند که آموزش مدل بر اساس لگاریتم درست‌نمایی، باعث می‌شود که مدل نتایج نامطلوب‌تری را تولید کند و به جای آن از معیار دیگری که در فصل ۲ بررسی شد استفاده کرددند. این معیار باعث شد نتایج بسیار مطلوبی از شبکه‌های پخشی به دست آید اما در ازای این که امتیاز کمی در معیار لگاریتم درست‌نمایی کسب کند. کار دومی که در این مقاله انجام شد این بود که با یادگیری واریانس‌هایی که در مقاله اصلی ثابت فرض شده بودند، توانستند تعداد تکرار^[۲۰] لازم برای رسیدن به نتیجه مطلوب را کاهش دهند و سرعت مدل را بسیار افزایش دهند. درنهایت در این مقاله نشان داده شد که کیفیت نمونه برداری و معیار لگاریتم درست‌نمایی با افزایش مقیاس^[۲۱] مدل، به طور نرم و یکنواختی تغییر می‌کند که این امر مقیاس‌پذیری^[۲۲] مدل را افزایش می‌دهد.

در بخش‌های بعد به هر کدام از این بهبودها و تاثیر آنها می‌پردازیم.

^[۱۹] Log Likelihood

^[۲۰] Iteration

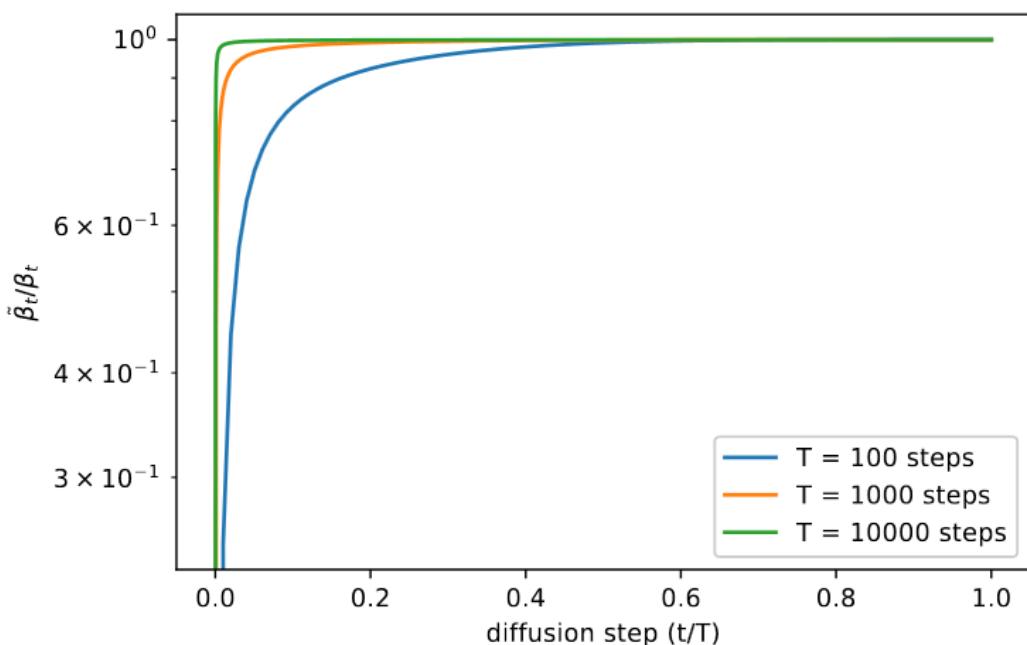
^[۲۱] Scale

^[۲۲] Scalability

۳-۱ بهبود مدل در معیار درست نمایی

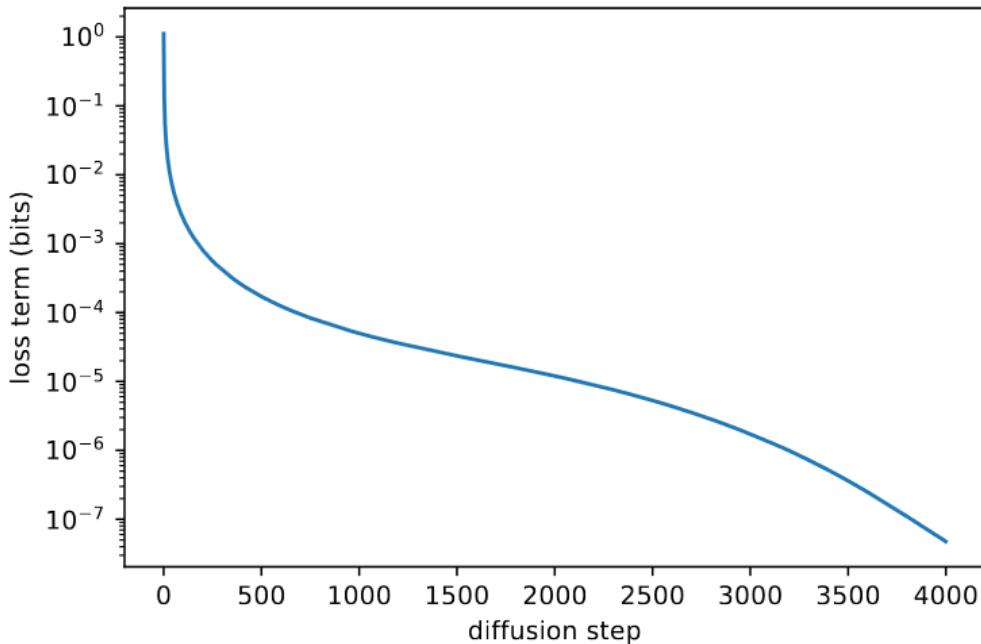
در مقاله مربوط به مدل‌های پخشی [۱] نویسنده‌گان علی‌رغم تولید نمونه‌های بسیار عالی، نتوانستند در معیار درست‌نمایی امتیاز بالایی بدست آورند. طبق مطالعاتی که در زمینه‌های مرتبط انجام شده بود، امتیاز درست‌نمایی بالا برای مدل‌های مولد بسیار بالاهمیت است زیرا نشان می‌دهد که یک مدل همه مودهای توزیع را به خوبی آموخته است. به این ترتیب در این مقاله [۲]، نویسنده‌گان به دنبال بهبود امتیاز درست‌نمایی مدل‌های پخشی رفته‌اند. اولین تغییری که در این جهت انجام شد افزایش مقدار T به عدد ۴۰۰۰ بود. نویسنده‌گان متوجه شدند با این افزایش مقدار لگاریتم درست‌نمایی از $3/99$ به $3/77$ کاهش پیدا می‌کند که پیشرفت قابل توجهی است.

یکی از فرض‌های اصلی و مهم مدل پخشی در مقاله مربوطه [۱] ثابت فرض کردن واریانس‌ها در مرحله مستقیم مدل بود. نمودار موجود در شکل ۱۰ نشان می‌دهد که نسبت واریانس ثابت و واریانس یادگرفته شده در تکرارهای زیاد به ۱ همگرا می‌شود و در واقع تفاوت بین این دو واریانس تنها در تکرارهای اولیه است که تصویر نهایی هنوز شکل نگرفته است. بنابراین فرض نویسنده‌گان مقاله اصلی [۱] خیلی اشتباه نبوده است.



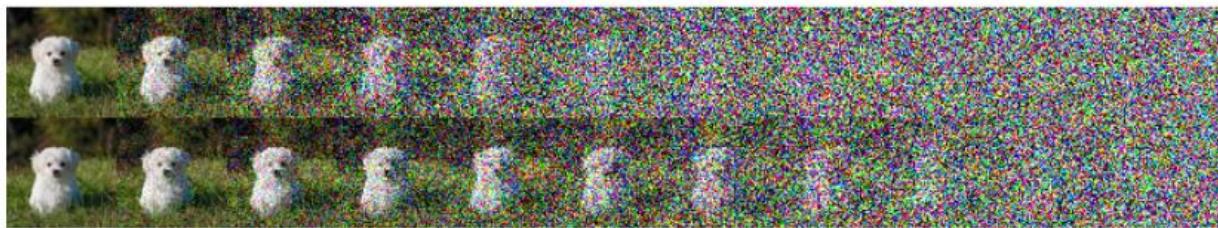
شکل ۱۰ - تغییر نسبت واریانس آموخته شده به واریانس ثابت در طی تکرارهای مختلف

اما اگر به نمودار موجود در شکل ۱۱ توجه کنید متوجه می‌شوید که همین تکرارهای اول هستند که بیشترین تاثیر را در مقدار درستنمایی دارند. بنابراین در مقاله [۲] با یادگیری مقدار واریانس‌ها در مرحله آموزش، امتیاز درستنمایی مدل‌های پخشی را نیز افزایش داده‌اند.



شکل ۱۱ - جملات تابع زیان نسبت به تکرار

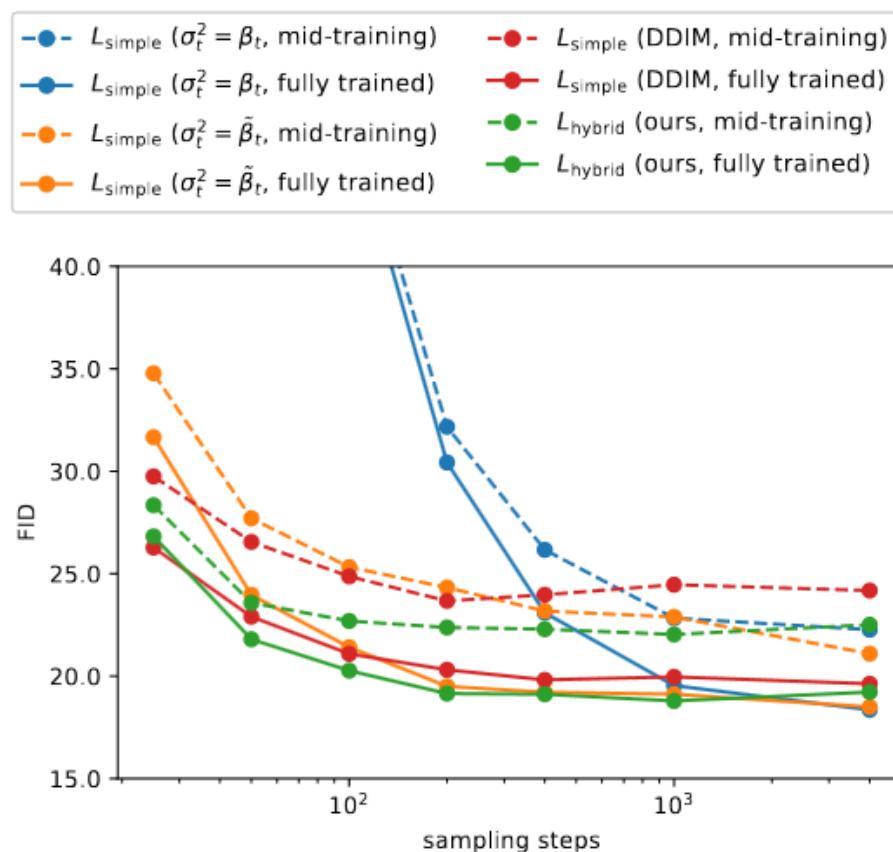
یکی دیگر از ایده‌های مهم این مقاله [۲] استفاده از زمان‌بند نویز کسینوسی به جای زمان‌بند نویز خطی است. با توجه به شکل ۱۲ می‌بینیم که هنگام استفاده از زمان‌بند خطی، در تعداد زیادی از مراحل داده موجود نویز خالص است که فایده زیادی نخواهد داشت، اما با استفاده از زمان‌بند کسینوسی، تعداد بیشتری از تکرارها دارای اطلاعات مفید از تصویر هستند.



شکل ۱۲ - زمان‌بند خطی (بالا) و زمان‌بند کسینوسی (پایین)

۲-۳ کاهش تعداد تکرارها

در حالت عادی، مدل‌های پخشی ای که با 4000 تکرار آموزش دیده‌اند، در زمان نمونه برداری نیز 4000 بار صدا زده می‌شوند. این کار بر روی پردازنده‌های گرافیکی امروزی به چندین دقیقه زمان نیاز دارد. در این مقاله^[۲] سعی شده با استفاده از تکنیک‌هایی، تعداد تکرارهای لازم در مرحله نمونه‌برداری کاهش پیدا کند. با توجه به آزمایش‌های انجام شده که نتایج آن‌ها در نمودار شکل ۱۳ مشاهده می‌شود، این کار برای مدل‌های پخشی ساده (رنگ آبی) امکان‌پذیر نیست زیرا معیار FID برای تصاویر تولید شده با تعداد تکرار کم بسیار نامناسب است. اگر به نموادرهای سبز رنگ در این شکل توجه کنید می‌بینید که برای این مدل‌ها که در این مقاله^[۲] معرفی شدند از این تکنیک نمونه برداری استفاده بکنیم، میزان امتیاز FID از تکرار 100 به بعد تقریباً ثابت می‌ماند و این امکان را می‌دهد که با سرعت بسیار بیشتری نمونه‌برداری را انجام دهیم.

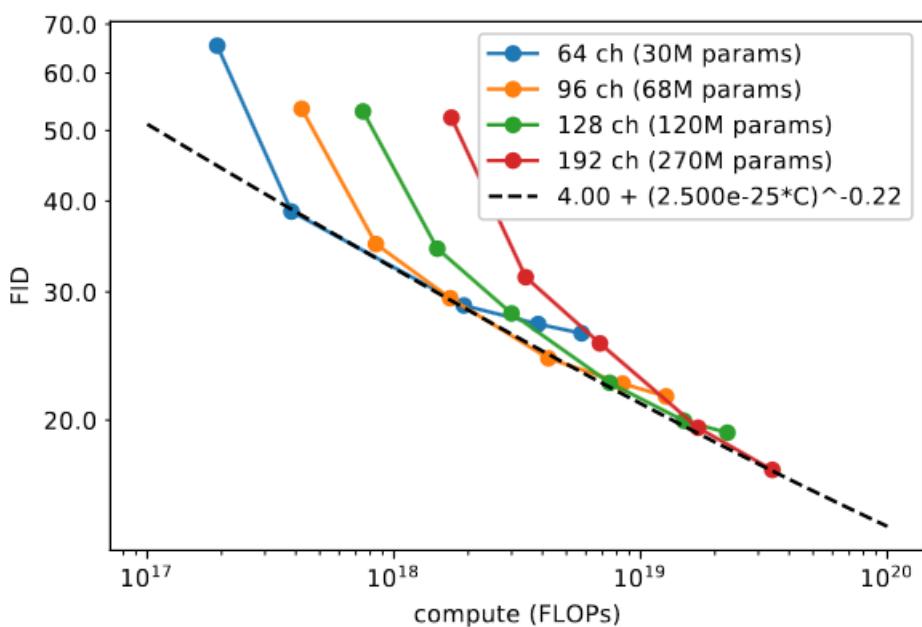


شکل ۱۳- نمونه برداری با تکرارهای مختلف

۳-۳ مقیاس پذیری مدل

با توجه به روند تکامل روش‌های یادگیری ماشین در طی سال‌های اخیر نویسندهان این مقاله^[۲] در نظر داشتن مقیاس‌پذیری مدل پخشی را مورد آزمایش قرار دهند و بیینند که آیا این مدل، با در دست داشتن داده بیشتر یا قدرت محاسباتی بیشتر قابل گسترش است یا خیر. مدل‌ها و معماری‌هایی که می‌توانند از تمام قدرت پردازشی موجود استفاده کنند مانند مدل‌های^[۲۳] در دنیای امروزی اهمیت بسیار زیادی دارند.

با توجه به نمودارهای موجود در شکل ۱۴ می‌توان دید که این مدل در معیار FID با افزایش قدرت پردازشی به خوبی رشد می‌کند.



شکل ۱۴ - بررسی مقیاس پذیری مدل پخشی

²³ Transformers

۴ کاربردهای مختلف مدل های پخشی

در سال های اخیر، مدل های پخشی توانسته اند توجه محققان زیادی را به خود جلب کنند. قدرت تولید نمونه، در کنار سرعت و خواص دیگری که این مدل ها دارند باعث شده به یکی از مباحث داغ روز تبدیل شوند و محققان مختلف سعی کنند از این مدل ها در زمینه های مختلف بهره برند.

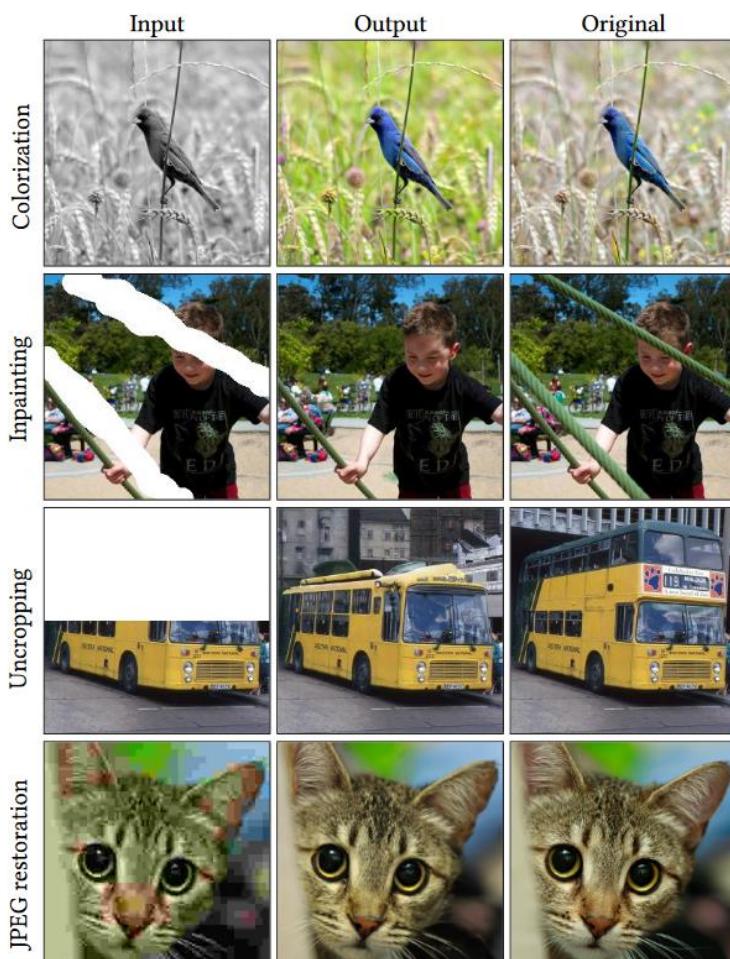


شکل ۱۵ - تعدادی از مثال های تبدیل متن به تصویر به کمک مدل 2 DALL-E

در این فصل تعدادی از کاربردهای مهم و معروف مدل های پخشی مانند DALL-E 2، Stable diffusion و DreamBooth که هر کدام از این مدل ها بسیار پرکاربرد و معروف هستند را بررسی خواهیم کرد. در این فصل به قدرت اصلی مدل های پخشی پی خواهیم برد و استفاده آنها در زمینه های مختلف مانند تبدیل متن به تصویر، تبدیل تصویر به تصویر، ویرایش خودکار تصویر، ایجاد تصاویر مشابه، بازسازی بخش های آسیب دیده تصویر و ... را نشان خواهیم داد.

Palette ۱-۴

در مقاله مربوط به مدل پالت^[۳] یک چارچوب یکسان برای انجام عملیات‌های مختلف تبدیل تصویر به تصویر^[۲۴] ارائه شده است. در این مقاله سعی شده از مدل‌های پخشی برای انجام کارهای مختلف مانند رنگی‌سازی تصاویر خاکستری^[۲۵]، بازسازی بخش‌هایی از تصویر^[۲۶]، گسترش تصویر^[۲۷] و رفع خودکار آسیب‌های بوجود آمده در اثر الگوریتم فشرده‌سازی JPEG^[۲۸] استفاده شود. این مقاله توانست در تمام این عملیات، به نتیجه بسیار عالی برسد که قدرت و کارایی مدل‌های پخشی را چندین برابر نشان می‌دهد.



شكل ۱۶ - کاربردهای مختلف روش پالت

²⁴ Image to Image

²⁵ Colorization

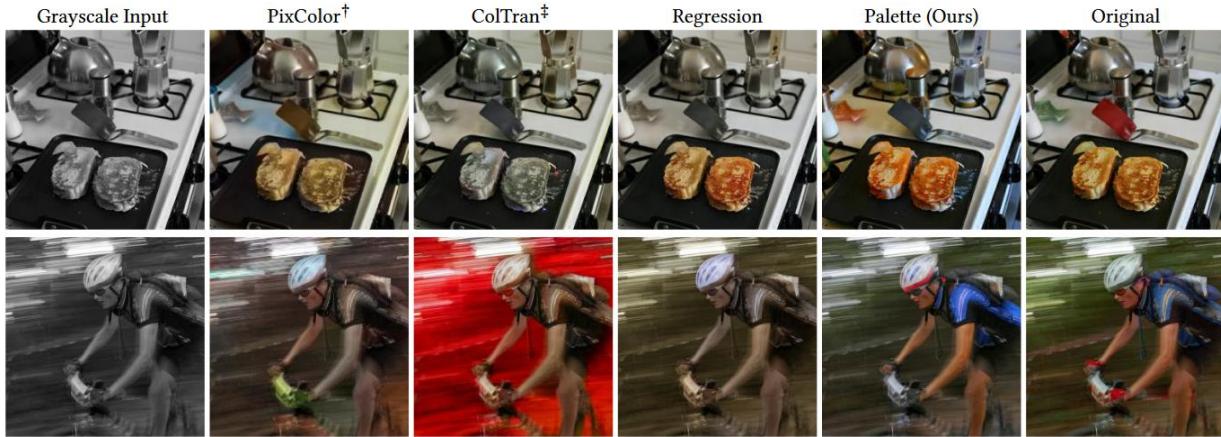
²⁶ Inpainting

²⁷ Uncropping

²⁸ JPEG Restoration

۱-۱-۴ رنگی سازی تصاویر خاکستری

یکی از مهم‌ترین مسائل موجود در زمینه بینایی ماشین رنگی کردن تصاویر سطح خاکستری است. در شکل ۱۷ مشاهده می‌کنیم که این روش از سایر روش‌های موجود در زمینه رنگی کردن تصاویر، نتیجه بهتری دارد. همچنین از نظر معیار-های FID و IS نیز این روش نسبت به سایر روش‌ها برتری دارد.



شکل ۱۷ - رنگی کردن تصاویر خاکستری

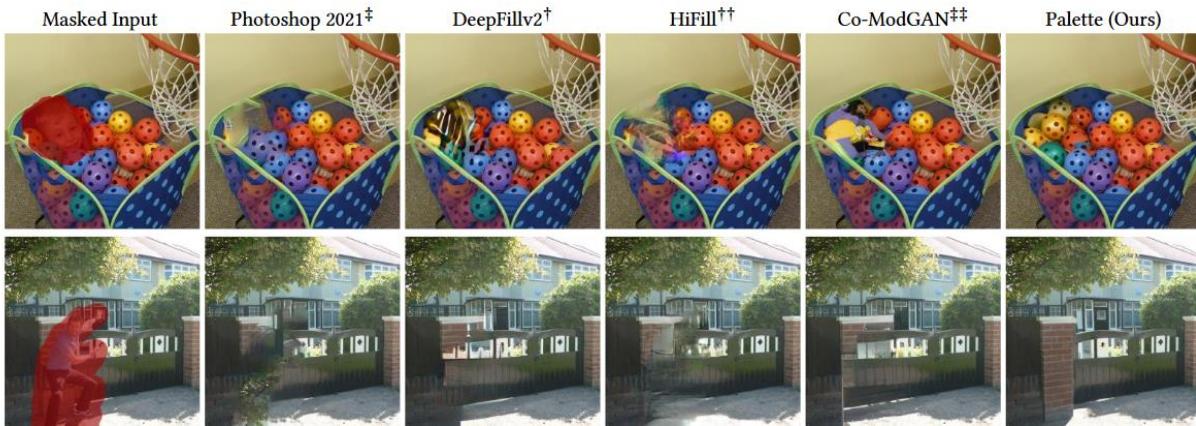
جدول ۲ - مقایسه روش پالت با سایر روش‌ها در زمینه رنگی کردن تصویر

Model	FID-5K ↓	IS ↑	CA ↑	PD ↓	Fool rate ↑
<i>Prior Work</i>					
pix2pix †	24.41	-	-	-	-
PixColor ‡	24.32	-	-	-	29.90%
Coltran ††	19.37	-	-	-	36.55%
<i>This paper</i>					
Regression	17.89	169.8	68.2%	60.0	39.45%
Palette	15.78	200.8	72.5%	46.2	47.80%
Original images	14.68	229.6	75.6%	0.0	-

۲-۱-۴ بازسازی تصویر

یکی دیگر از کاربردهای رایج و پراهمیت در زمینه پردازش تصویر، عمل درونیابی بخش‌هایی از تصویر است. ترمیم بخش‌هایی از تصاویر اسکن شده، حذف کردن اجسامی از تصویر، حذف موادی بین دوربین و جسم مورد نظر و ... کاربردهای مختلف این روش هستند.

در شکل ۱۸ می‌توانید مقایسه‌ای از روش‌های مختلف را در این مسئله مشاهده کنید. همان‌طور که دیده می‌شود، مدل پخشی در این زمینه نیز بهترین نتیجه را داده است.



شکل ۱۸ - مقایسه روش‌های مختلف در مسئله ترمیم تصویر

۱-۴-۳- گسترش تصویر

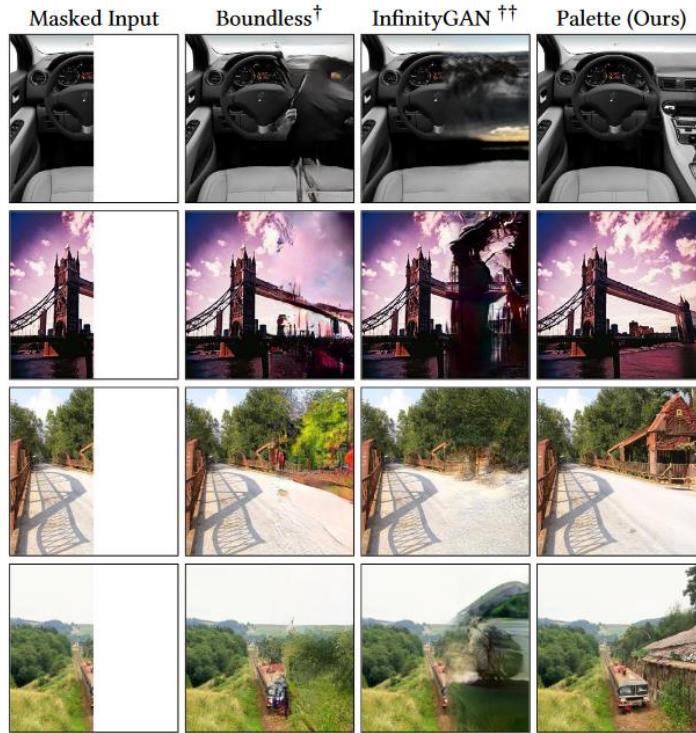
یکی از عملیات چالش برانگیز موجود دیگر، عمل گسترش تصویر است. این عملیات از ترمیم یا درون‌یابی تصویر دشوارتر است زیرا اطلاعات محلی اطراف ناحیه‌هایی که باید تولید شوند به اندازه روش ترمیم تصویر وجود ندارد و نتایج مدل رندوم تر خواهد بود. در این مقاله برای تبدیل تصاویر معمولی به پاناروما^{۲۹}، ابتدا یک ناحیه مربعی وسط پاناروما به مدل داده می‌شود و بخش کوچکی از دو طرف تصویر گسترش می‌یابد. تصویر حاصل به عنوان تصویر ورودی مرحله بعد در نظر گرفته می‌شود و این عمل تا زمانی که به اندازه دلخواه برسیم تکرار می‌شود. نتیجه تصاویر پانارومای بی‌نظیر است که در شکل ۱۹ مشاهده می‌کنید.



شکل ۱۹ - نتیجه گسترش تصویر به کمک روش پالت

²⁹ Panorama

گسترش تصویر می‌تواند کاربردهای متفاوتی از تولید پاناروما نیز داشته باشد. هنگامی که یک تصویر بریده^{۳۰} شده باشد و بخش‌های بریده شده تصویر موردنیاز باشند نیز، گسترش تصویر می‌تواند گزینه مناسبی باشد. همان‌طور که در تصویر ۲۰ قابل مشاهده است، روش پالت از روش‌های رقیب نتیجه بهتری دارد.



شکل ۲۰ - مقایسه نتایج گسترش تصویر به کمک روش پالت و سایر روش‌ها

۴-۱-۴ رفع آسیب‌های JPEG

پرکاربردترین الگوریتم فشرده‌سازی تصویر در حال حاضر الگوریتم JPEG است. این الگوریتم یک ساختار بلاکی دارد که در اثر آن، مخصوصاً وقتی نسبت فشرده سازی بالایی استفاده می‌شود باعث می‌شود تعدادی بلاک در تصویر مشاهده شود. رفع این آسیب‌ها اهمیت بسیار زیادی دارد زیرا این امکان را به ما می‌دهد که از تنظیمات فشرده‌سازی با نرخ فشرده‌سازی بیشتری استفاده کنیم و به هنگام نیاز تصاویر، آن‌ها را با جزئیات زیاد بازسازی کنیم.

³⁰ Crop



شکل ۲۱ - نتیجه کاهش آسیب های JPEG

همان‌طور که در تصویر ۲۱ دیده می‌شود، تصاویری که در اثر فشرده‌سازی زیاد (QF=5) آسیب دیده‌اند، توسط این روش به خوبی بازیابی می‌شوند. این روش جزیيات تصویر را نیز به خوبی بازسازی می‌کند و تصاویر مات خروجی نمی‌دهد.

۴-۱-۵ نتیجه گیری

با توجه به کاربردهایی که در این بخش دیدیم، می‌توان با اعتماد به نفس بیشتری به مدل‌های پخشی اطمینان کرد و از آن‌ها، بدون ایجاد تغییرات زیاد، در کاربردهایی که هنوز به نتایج مطلوب نرسیده‌اند استفاده کرد. این مدل‌ها در طی سال‌های اخیر قدرت تولید تصویر خود را به خوبی به همگان نشان داده‌اند.

DALL-E ۲-۴

در سال ۲۰۲۱، کمپانی OpenAI، به دنبال موفقیت‌های معماری مبدل^{۳۱}، روش تبدیل متن به تصویر DALL-E منتشر کرد. حدود یک سال بعد و پس از دیدن موفقیت مدل‌های پخشنی، در مقاله Hierarchical Text-Conditional Image Generation with CLIP Latents [۴] دال ای ۲ را منتشر کرد که در دو بخش اصلی آن از مدل‌های پخشنی استفاده شده بود.



a shiba inu wearing a beret and black turtleneck

شکل ۲۲ - نمونه تصویر تولید شده توسط دال ای ۲ به همراه متن ورودی

مدل دال ای ۲ به طور کلی از سه بخش زیر تشکیل شده است:

- یک واحد تبدیل متن به بردار با استفاده از CLIP^{۳۲}
- یک شبکه عصبی پیشین^{۳۳} به کمک مدل‌های پخشنی
- یک شبکه کدگشا^{۳۴} به کمک مدل‌های پخشنی

شکل ۲۳ معماری کلی روش دال ای ۲ را نشان می‌دهد. نحوه کلی عملکرد این مدل به این شکل است که یک کدگذار^{۳۵} تصویر و یک کدگذار برچسب آن تصویر، به طور همزمان به کمک معیار CLIP آموزش می‌بینند. این معیار یک معیار تضادی است. نحوه آموزش توسط این معیار به این شکل است که دو تصویر با برچسب مشابه به مدل داده می‌شود و مدل یک امتیاز شباهت برای این دو ورودی تولید می‌کند. پس از اتمام آموزش انتظار می‌رود مدل بتواند برای تصاویر مشابه و برچسب‌های مشابه، بردارهای کدشده‌ی یکسانی تولید کند.

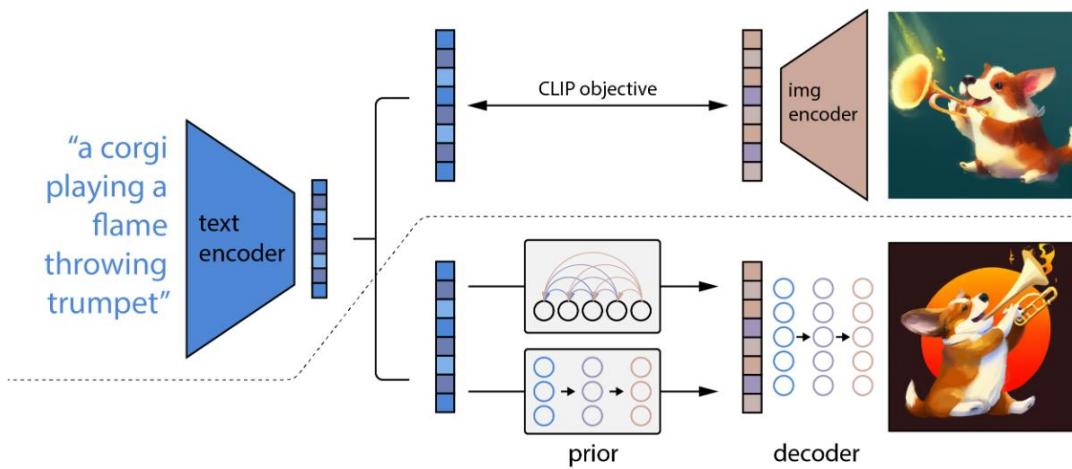
³¹ Transformer

³² Contrastive Language-Image Pre-Training

³³ Prior

³⁴ Decoder

³⁵ Encoder



شکل ۲۳ - شمای کلی معماری دال ای ۲

پس از انجام این مرحله، کدگذارهای ذکر شده قفل^{۳۶} می‌شوند. بخشی که در شکل ۲۳ زیر نقطه چین رسم شده است، بخش تولید تصویر جدید است. این بخش ابتدا برای هر ورودی متن، یک بردار متن(بردار آبی رنگ) می‌گیرد. این بردار به عنوان ورودی به مدل پیشین داده می‌شود و به عنوان خروجی، یک بردار تصویر(بردار نارنجی رنگ) مشابه آنچه توسط کدگذار بخش قبل تولید شده بود خروجی داده می‌شود. این بردار در هنگام آموزش در اختیار مدل پیشین است اما پس از اتمام آموزش انتظار می‌رود مدل پیشین بر اساس اطلاعاتی که صرفاً از بردار متنی می‌گیرد بتواند این بردار تصویری را تولید کند. دلیل استفاده از معیار CLIP برای آموزش دو کدگذار بالا و آموزش همزمان آنها همین است. پس از این که بردار تصویر موردنظر تولید شد، به عنوان ورودی به یک مدل پخشی داده می‌شود و مانند آنچه در بخش‌های قبلی این گزارش دیده‌ایم، تصویر مدنظر تولید می‌شود.

نکته قابل توجه این است که ما لزوماً به مدل پیشین نیازی نداریم و می‌توانیم بردار متنی(آبی) یا حتی متن خام را به مدل کدگشای مرحله آخر بدهیم اما اینکار باعث کاهش کیفیت و باورپذیری تصاویر تولید شده می‌شود. اهمیت وجود مدل پیشین را در ادامه بررسی خواهیم کرد.

³⁶ Freeze

۱-۲-۴ جزیات روش تولید تصاویر

با در نظر گرفتن یک جفت داده ورودی (y, x)، که در آن x تصویر و y متن مربوط به آن است، توقع داریم یک مدل مولد آماری به شکل $P(x|y)$ بسازیم که با داشتن بردار متغیر y ، بتواند تصویر x را تولید کند. در این روش همان‌طور که ذکر شد، یک بردار میانی z وجود دارد که به کمک CLIP تولید می‌شود. ابتدا مدل پیشین $(P(z|y))$ با دیدن متن ورودی، بردار میانی z را تولید می‌کند و سپس کدگشا، $P(x|z, y)$ با دیدن این بردار میانی، تصویر x را تولید می‌کند. در این بخش می‌توان به طور دلخواه متن خام y را نیز به مدل کدگشا داد. درواقع به طور کلی طبق رابطه ۳ داریم:

$$P(x|y) = P(x, z|y) = P(x|z, y)P(z|y) \quad (3)$$

دلیل برقرار بودن تساوی اول این است که z از روی x بدست می‌آید و به کمک قانون زنجیره‌ای به تساوی دوم می‌رسیم.

همچنین، یکی از مشکلات روش‌های پخشی این است که سرعت آموزش و اجرای آنها وابسته به اندازه تصویر ورودی است و در حال حاضر برای تصاویر بزرگتر از ۶۴ در ۶۴ پیکسل بسیار کند هستند. اما خوشبختانه نیازی نیست در طی عمل رفع نویز، تمام پیکسل‌های تصویر را تولید کنیم و همینکه یک شمای کلی از تصویر مورد نظر بدست بیاوریم کافی است. در این مقاله و بسیاری از مقالات دیگر پس از ایجاد تصویر ۶۴ در ۶۴ توسط مدل پخشی، با دوبار اعمال یک مدل بزرگ‌نمایی^{۳۷}، می‌توان به اندازه ۱۰۲۴ در ۱۰۲۴ بدون کاهش کیفیت رسید زیرا این مدل‌های افزایش اندازه در حال بزرگ‌نمایی تصویر، به آن جزیيات نیز اضافه می‌کنند زیرا به صورت توأم با مدل پخشی آموزش دیده اند.

۲-۲-۴ نتایج و ویژگی‌ها

اگر به شکل ۲۴ نگاه کنید متجه قدرت تولید تصاویر به کمک بردار معنایی ساخته شده توسط CLIP خواهد شد. تصاویر تولید شده در هر دسته، از لحاظ معنایی بسیار شبیه هستند به عنوان مثال در تصاویر سمت چپ، همه تصاویر یک تصویر ساحل دارای ساعت و تنہ درخت هستند اما جزیيات بی‌اهمیت آنها دچار تغییر شده است. یا در تصویر راست، پس زمینه تمامی تصاویر مشابه است و تعدادی خط سفید روی هم در همه تصاویر تولید شده قرار دارد.

³⁷ Upscaling



شکل ۲۴- تولید تصاویر مشابه از نظر معنایی به کمک بردار معنایی CLIP



شکل ۲۵ - تعدادی از تصاویر تولیدشده توسط دال ای ۲

جدول ۳ - مقایسه عملکرد روش دال ای ۲ و سایر روش‌ها

Model	FID	Zero-shot FID	Zero-shot FID (filt)
AttnGAN (Xu et al., 2017)	35.49		
DM-GAN (Zhu et al., 2019)	32.64		
DF-GAN (Tao et al., 2020)	21.42		
DM-GAN + CL (Ye et al., 2021)	20.79		
XMC-GAN (Zhang et al., 2021)	9.33		
LAFITE (Zhou et al., 2021)	8.12		
Make-A-Scene (Gafni et al., 2022)	7.55		
DALL-E (Ramesh et al., 2021)	~ 28		
LAFITE (Zhou et al., 2021)	26.94		
GLIDE (Nichol et al., 2021)	12.24	12.89	
Make-A-Scene (Gafni et al., 2022)		11.84	
unCLIP (AR prior)	10.63	11.08	
unCLIP (Diffusion prior)	10.39	10.87	

با توجه به تصاویر تولیدشده و نتایج عددی، دال ای ۲ در زمان انتشار یکی از برترین و پرسر و صدا ترین مدل‌های تبدیل متن به تصویر بود.

۴-۲-۳ اهمیت مدل پیشین

در بخش‌های قبل در مورد چگونگی عملکرد مدل پیشین بحث کردیم. در این بخش خواهیم دید که این مدل پیشین تا چه حد اهمیت دارد و چه میزان در کیفیت تصاویر تولید شده موثر است.



شكل بالا به خوبی این اهمیت وجود این مدل را نشان می‌دهد، ردیف اول تصاویر تولید شده با استفاده از ورودی متن خام هستند. تصاویر سطر دوم به جای متن خام، از بردار متن تولید شده به کمک CLIP استفاده می‌کنند و تصاویر سطر سوم از بردارهای تصویر تولید شده به کمک مدل پیشین استفاده کرده اند. با مقایسه متن ورودی و تصاویر تولید شده به خوبی می‌توان اهمیت مدل پیشین را مشاهده کرد. تصاویری که از این بردار تصویر استفاده کرده اند به خوبی تمام مواردی که در متن گفته شده است را دارا هستند.

۴-۲-۴ نتیجه گیری

مدل دال ای ۲ یک سال پس از انتشار همچنان یکی از بهترین مدل‌های تبدیل متن به تصویر است. عملکرد بسیار خوب این مدل بار دیگر کارا بودن مدل‌های پخشی را در تولید تصویر گواهی می‌کند. این مدل در حال حاضر به صورت متن باز در اختیار عموم قرار داده نشده و به صورت یک پایانه اینترنتی می‌توان از آن استفاده کرد. طبق گفته‌های سازنده‌های دال ای ۲، قابلیت تولید تصاویر خلاف قانون توسط دال ای ۲، یکی از دلایلی است که به صورت رایگان و عمومی منتشر نشده است. باید دید در آینده این قوانین چه تغییری خواهند کرد.

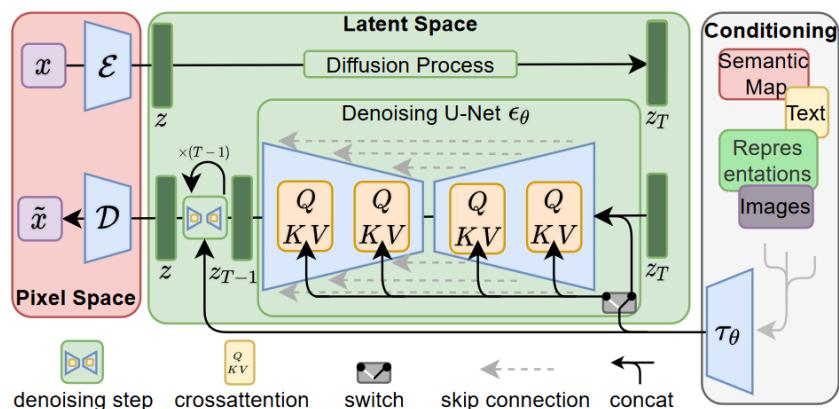
Stable Diffusion ۳-۴

یکی از معروف‌ترین و جدیدترین مدل‌های تولید تصویر که اخیراً نیز منتشر شده از مقاله [5] Synthesis with Latent Diffusion Models سرچشمه می‌گیرد. این مدل در کنار مدل 2 DALL-E دو مدلی هستند که با نتایج خیره‌کننده خود، توجه عموم، سرمایه‌گذاران و محققان سرتاسر جهان را به مدل‌های مولد تصویر و به خصوصی مدل‌های پخشی جلب کرده‌اند. از نظر بسیاری از مردم، این مدل‌ها یک انقلاب بزرگ در صنعت هوش مصنوعی به شمار می‌آیند.

همان‌طور که قبل‌تر نیز اشاره شد، یکی از مشکلات اصلی مدل‌های پخشی سرعت پایین انجام مرحله رفع‌نویز که یک مرحله تکراری است می‌باشد. در این مقاله [5] برای بهبود این مشکل این ایده مطرح شد که عمل رفع‌نویز را به جای اینکه در فضای پیکسلی انجام دهیم بهتر است در یک فضای نهان میانی با ابعاد کمتر انجام دهیم. لازم به ذکر است که مدل‌های پخشی قادر هستند به طور خودکار اطلاعات غیرمهم در فضای پیکسلی را نادیده بگیرند اما با کارکردن در فضای پیکسلی همچنان لازم است مقدار زیادی محاسبات بی‌فایده بر روی پیکسل‌ها انجام شود. با رفتتن به فضای میانی و انجام عملیات طولانی و تکراری رفع‌نویز در آن فضا، این مدل‌ها به حد بهینگی بسیار بالایی خواهند رسید.

۴-۳-۱ معماری و روش

به طور کلی در این مقاله [5]، یک روش دو مرحله‌ای که شامل یک مدل پخشی که بر روی بردارهای میانی عمل می‌کند و یک سیستم کدگذار-کدگشا، که تصاویر را به بردار میانی و از بردار میانی رفع‌نویز شده توسط مدل پخشی به تصویر نهایی تبدیل می‌کند است، ارائه شده است.

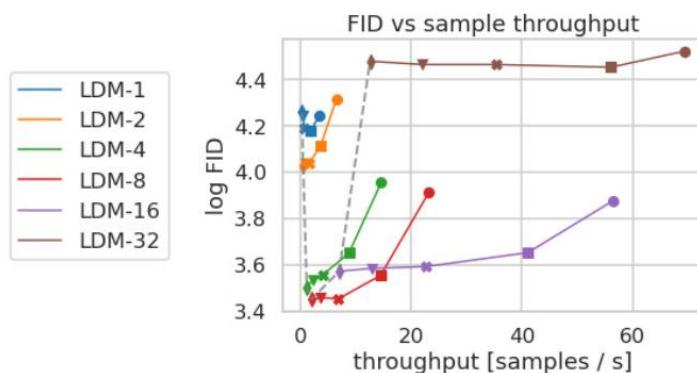


شکل ۲۶- معماری کلی روش LDM³⁸

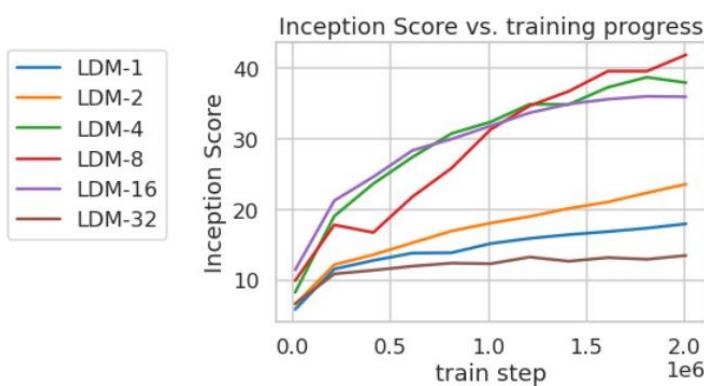
³⁸ Latent Diffusion Models با نام تجاری Stable Diffusion

در شکل ۲۶ به خوبی نشان داده است که تصویر ورودی x پس از عبور از کدگذار، به بردار میانی با ابعاد کوچک-تر Z تبدیل می‌شود. عمل نویزی‌شدن بر روی این بردار انجام می‌شود(به جای نویزی کردن تصویر x ، سپس این بردار میانی به همراه ورودی‌های شرطی دیگر مانند متن ورودی و، به شبکه عصبی کاهش نویز که در این مقاله یک شبکه UNET است داده می‌شود. این اتفاق T بار تکرار می‌شود تا به بردار Z بدون نویز برسیم. در نهایت توسط واحد کدگشا بردار Z به تصویر تولیدی تبدیل می‌شود. یکی از تغییراتی که در شبکه عصبی این مدل نسبت به مدل اصلی در مقاله مدل‌های پخشی^[۱] ایجاد شده است استفاده از ساختار توجه مشابه معماری مبدل‌ها است. طبق آزمایش‌های نویسنندگان این مقاله^[۵]، استفاده از این مکانیزم باعث ایجاد تصاویر بهتری شده است.

با توجه به شکل ۲۷ و ۲۸، می‌توان تاثیر کاربروی ابعاد کوچک‌تر در مرحله پخش را مشاهده کرد. منحنی آبی در شکل نشان دهنده مدلی است که هیچ کاهش بعدی ندارد و بر روی پیکسل‌ها کار می‌کند. منحنی قهوه‌ای مدلی است که تصویر ورودی را ۳۲ برابر کوچک‌تر می‌کند. آزمایش‌ها نشان می‌دهد که کاهش ابعاد تصویر به میزان ۴ تا ۱۶ برابر بهترین نتیجه را می‌دهد و نمودارهای شکل ۲۷ و ۲۸ نیز این نظریه را اثبات می‌کنند.



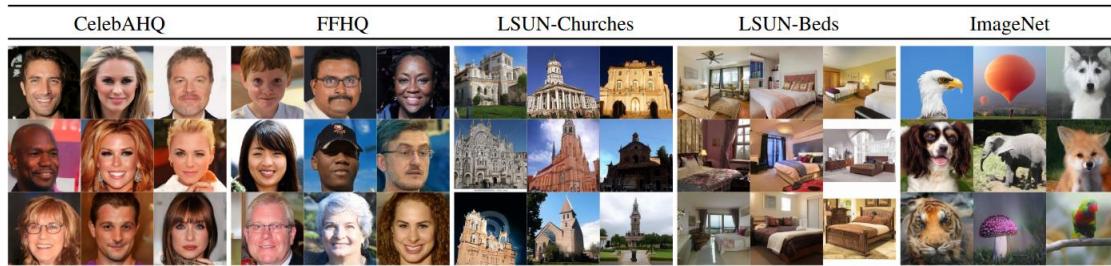
شکل ۲۷- مقایسه اثر میزان کاهش ابعاد بر معیار FID (کمتر=بهتر)



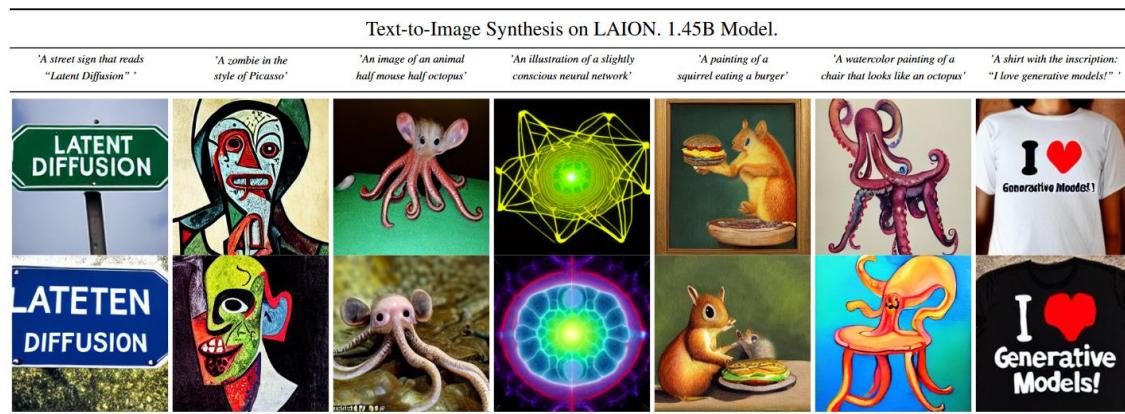
شکل ۲۸ - مقایسه اثر میزان کاهش ابعاد بر معیار IS بر حسب میزان آموزش (بیشتر=بهتر)

۴-۳-۲ نتایج و ویژگی ها

در شکل ۲۹ می توان تعدادی از تصاویر تولید شده توسط این روش را مشاهده کرد. تصاویر تولید شده توسط این روش بسیار عالی و باورپذیر هستند.



شکل ۲۹ - برخی از تصاویر تولید شده توسط روش LDM

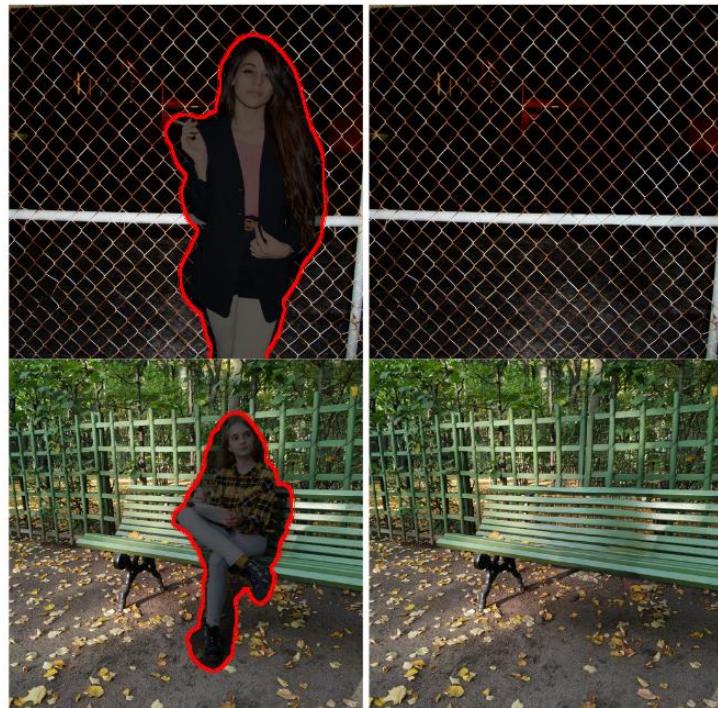


شکل ۳۰ - برخی از تصاویر تولید شده با ورودی متن توسط روش LDM

این مدل در تولید تصاویر بدون شرط به بهترین نتایج روز رسیده است و در مسئله تولید تصویر به کمک متن ورودی نیز به نتایج بسیار رقابتی ای رسیده است. در شکل ۳۰ می توان برخی از تصاویر تولید شده توسط متن، به همراه متن داده شده را مشاهده کرد. در جدول ۴ و ۵، نتایج این روش با سایر روش‌ها مقایسه شده است که برتری این مدل را اثبات می‌کند.

در فصل‌های قبل دیدیم که تولید تصاویر از نویز خالص، تنها کاربرد مدل‌های پیش‌بینی نیست. ایده انجام عملیات بر روی بردار میانی نیز صرفا برای تولید تصاویر نیست، همان‌طور که انتظار می‌رود از مدل‌های LDM می‌توان برای سایر مسائل مشابه که مدل‌های پیش‌بینی در آن‌ها عملکرد خوبی دارند نیز استفاده کرد.

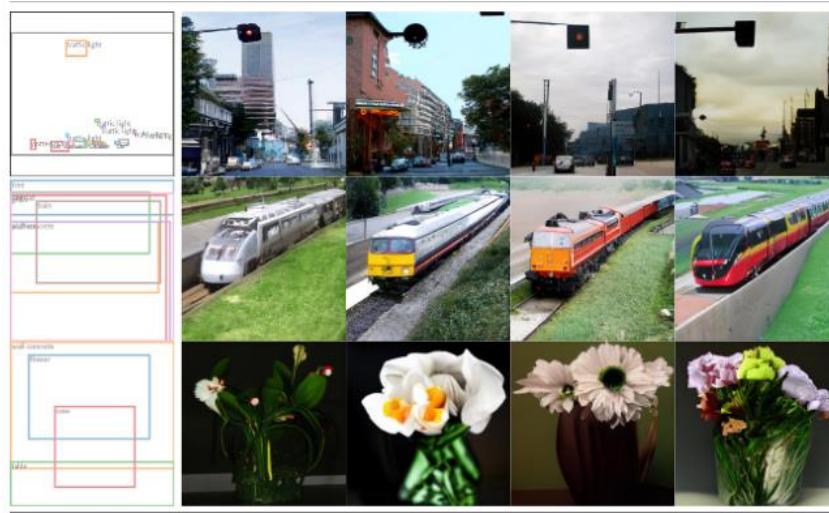
در شکل ۳۱ می‌توان عملکرد مدل‌های LDM در مسئله ترمیم تصویر را مشاهده کرد. عملکرد دقیق این مدل در تمام مسائلی که در ادامه بحث می‌شود در مقاله اصلی قابل مشاهده است. در شکل ۳۲ یکی از نمونه تصاویر تولید شده با مشخص کردن برچسب معنایی هر پیکسل قابل مشاهده است و شکل ۳۳ مربوط به تولید تصویر با ورودی مستطیل‌هایی با کلاس‌های مختلف است.



شکل ۳۱ - نتیجه ترمیم تصویر به کمک روش LDM



شکل ۳۲ - نتیجه تولید تصویر بر اساس برچسب معنایی پیکسل ها



شکل ۳۳ - نتیجه تولید تصویر به کمک برچسب کلاس نواحی ای از تصویر

جدول ۴ - مقایسه روش LDM و سایر روش‌ها در تولید تصویر بدون شرط

CelebA-HQ 256 × 256				FFHQ 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [63]	15.8	-	-	ImageBART [21]	9.57	-	-
VQGAN+T. [23] (k=400)	10.2	-	-	U-Net GAN (+aug) [77]	10.9 (7.6)	-	-
PGGAN [39]	8.0	-	-	UDM [43]	5.54	-	-
LSGM [93]	7.22	-	-	StyleGAN [41]	4.16	0.71	0.46
UDM [43]	7.16	-	-	ProjectedGAN [76]	3.08	0.65	0.46
<i>LDM-4</i> (ours, 500-s [†])	5.11	0.72	0.49	<i>LDM-4</i> (ours, 200-s)	4.98	0.73	0.50
LSUN-Churches 256 × 256				LSUN-Bedrooms 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑	Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-	ImageBART [21]	5.51	-	-
ImageBART [21]	7.32	-	-	DDPM [30]	4.9	-	-
PGGAN [39]	6.42	-	-	UDM [43]	4.57	-	-
StyleGAN [41]	4.21	-	-	StyleGAN [41]	2.35	0.59	0.48
StyleGAN2 [42]	3.86	-	-	ADM [15]	1.90	0.66	0.51
ProjectedGAN [76]	1.59	<u>0.61</u>	<u>0.44</u>	ProjectedGAN [76]	1.52	<u>0.61</u>	0.34
<i>LDM-8*</i> (ours, 200-s)	4.02	0.64	0.52	<i>LDM-4</i> (ours, 200-s)	2.95	0.66	<u>0.48</u>

جدول ۵ - مقایسه روش LDM و سایر روش‌ها در تولید تصویر با ورودی متن

Text-Conditional Image Synthesis				
Method	FID ↓	IS↑	Nparams	
CogView [†] [17]	27.10	18.20	4B	self-ranking, rejection rate 0.017
LAFITE [†] [109]	26.94	<u>26.02</u>	75M	
GLIDE* [59]	12.24	-	6B	277 DDIM steps, c.f.g. [32] $s = 3$
Make-A-Scene* [26]	11.84	-	4B	c.f.g for AR models [98] $s = 5$
<i>LDM-KL-8</i>	23.31	20.03 ± 0.33	1.45B	250 DDIM steps
<i>LDM-KL-8-G*</i>	12.63	30.29 ± 0.42	1.45B	250 DDIM steps, c.f.g. [32] $s = 1.5$

Dreambooth ۴-۴

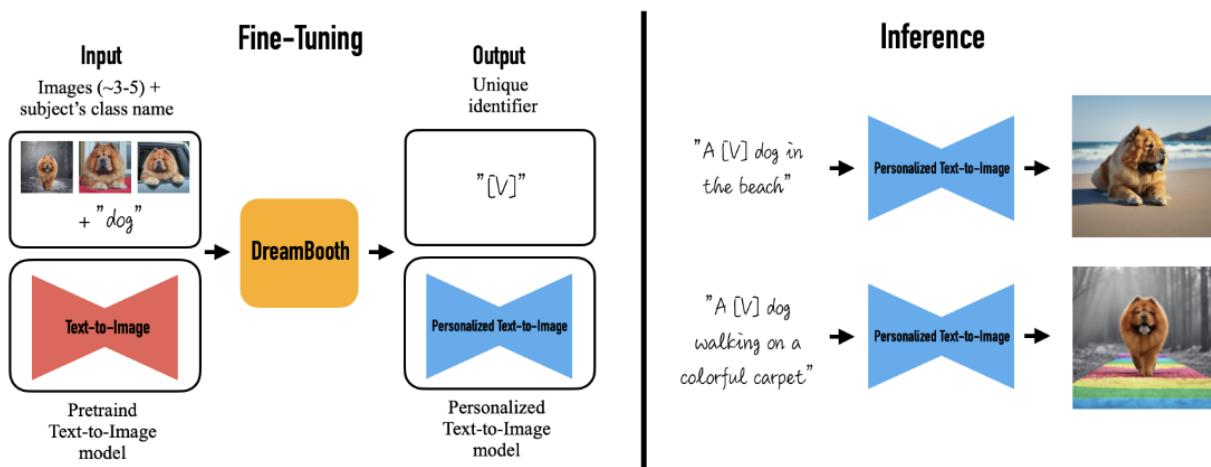
آخرین مقاله‌ای که در این پژوهه مورد بحث قرار می‌گیرد مقاله‌ای تحت عنوان DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation است. یکی از اصلی‌ترین مشکلات مدل‌های تبدیل متن به تصویر افزودن اطلاعات جدید به مدل است. اکثر مدل‌های زبانی مورد استفاده در سیستم‌های تبدیل متن به تصویر، بر روی داده‌های متنوع و وسیعی که در سطح اینترنت در دسترس است آموزش دیده‌اند. به این خاطر است که این مدل‌ها در تولید اشیا یا حیوانات یا مفاهیم کلی هیچ گونه مشکلی ندارند زیرا احتمالاً چندین و چندبار با این مفاهیم در داده‌های آموزشی خود روبرو شده‌اند. سوالی که پیش می‌آید این است که اگر بخواهیم به مدل بگوییم تصویری از خودمان تولید کند چه؟ مدل از کجا باید بداند ما چه شکلی هستیم؟ جواب این سوال‌ها در این مقاله [۶] به خوبی داده شده‌اند. روش DreamBooth همان‌طور که از اسمش نیز پیدا است، مانند یک اتاقک عکاسی است که مشابه رویا دیدن تصاویر جدیدی تولید می‌کند. این روش به صورت few-shot عمل می‌کند و با گرفتن ۳ الی ۵ تصویر نمونه و در نظر یک شناسه برای شخص یا جسم موردنظر، می‌آموزد که از این پس در صورت نیاز به چه صورت، شی یا فرد داده شده را تولید کند.



شکل ۳۴ - تصاویر تولیدشده شخصی سازی شده با کمک روش DreamBooth

۱-۴-۴ معماری و روش

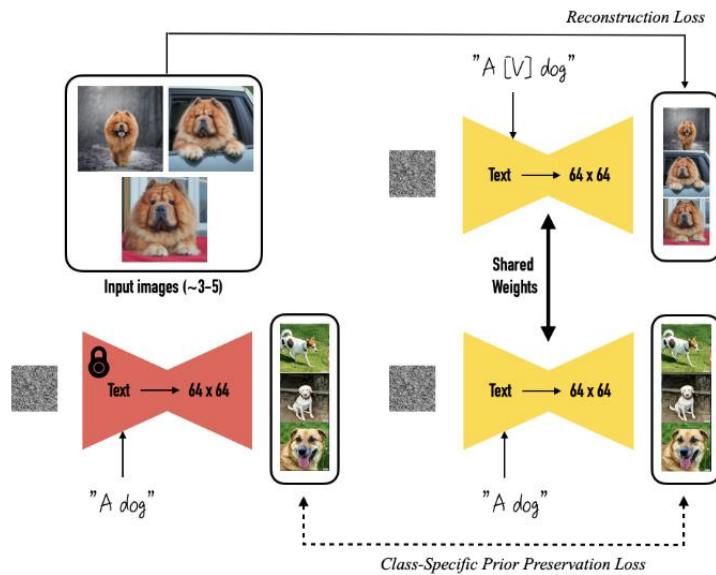
با توجه به شکل ۳۵ که شمای کلی سیستم را نشان می‌دهد، متوجه می‌شویم که این روش، در واقع یک مدل تبدیل متن به تصویر از پیش آموزش دیده را بهینه^{۳۹} می‌کند. ساختار قرمز رنگ در شکل، یک مدل تبدیل متن به تصویر از پیش آموزش دیده شده است. روش DreamBooth با گرفتن این مدل و تعدادی تصویر نمونه از سوژه، یک مدل بهینه شده



شکل ۳۵ - شمای کلی سیستم DreamBooth

برای تبدیل متن به تصویر و یک شناسه مخصوص سوژه ارائه می‌دهد. در زمان اجرا هرگاه مدل در متن ورودی، شناسه تولیدشده برای سوژه V را ببیند، در تصویر تولید شده اطلاعاتی که از تصاویر نمونه مربوط به V آموخته است را پیاده می‌کند.

³⁹ Finetune



شکل ۳۶ - نحوه آموزش مدل DreamBooth

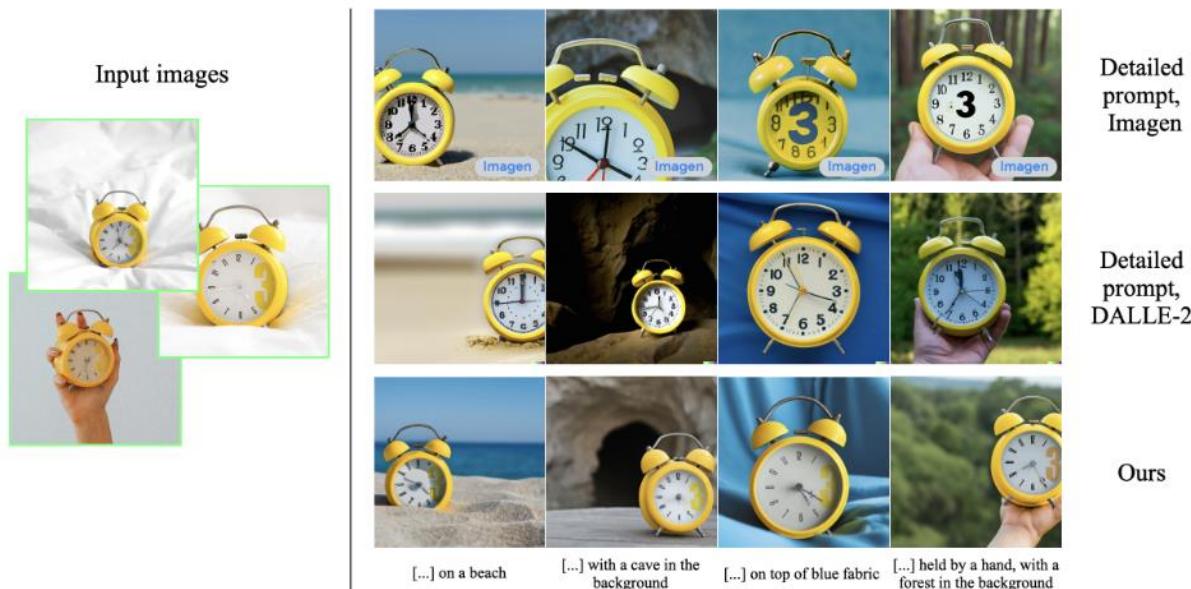
شکل ۳۶ نحوه آموزش این مدل را بهتر نشان می‌دهد. به یک مدل تبدیل متن به تصویر، به طور همزمان متن‌های "A" و "A [v]" و "A [class]" داده می‌شوند و مدل به طور توأم با استفاده از دوتابع زیان بازسازی تصاویر نمونه^{۴۰} و تابع زیان نگاهداری کلاس پیشین^{۴۱} آموزش می‌بیند. پس از اتمام آموزش، مدل زرد رنگ می‌تواند از توکن "[v]" در تولید تصاویر به خوبی استفاده کند.

⁴⁰ Reconstruction Loss

⁴¹ Class-specific Prior Preservation Loss

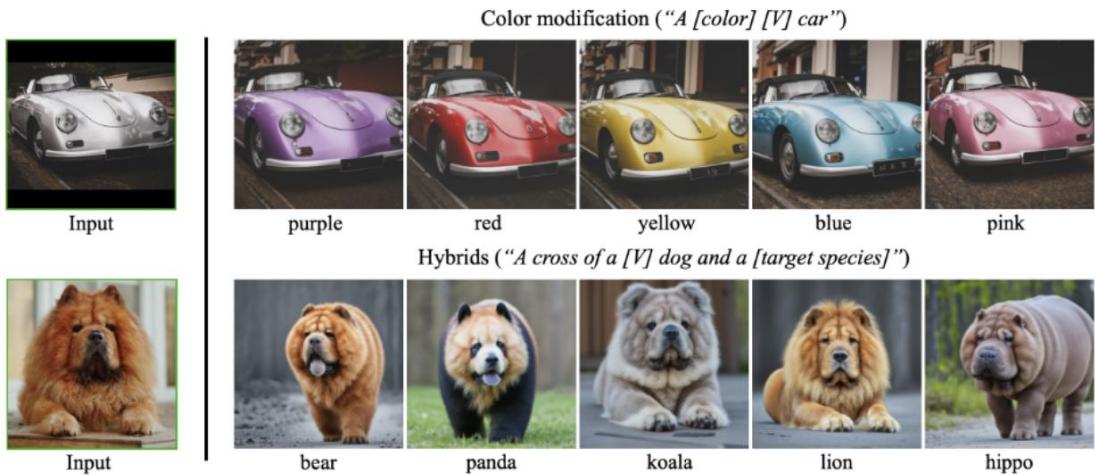
۴-۱ نتایج و کاربردها

همان‌طور که گفته شد، پیش از این، روش‌های تبدیل متن به تصویر، در یادگیری اطلاعات جدید مشکل داشتند و تمام تصاویری که می‌توانستند تولید کنند بر اساس دانشی بود که هنگام آموزش و با استفاده از مجموعه‌داده آموزشی کسب کرده بودند بود. بنابراین مدل DreamBooth در هر زمینه‌ای که نیاز باشد یک سوژه خاص را در تصویر خروجی به هر شکلی داشته باشیم بسیار بهتر از رقبا عمل می‌کند.



شکل ۳۷ - مقایسه روش‌های مختلف تولید تصویر با سوژه مشخص

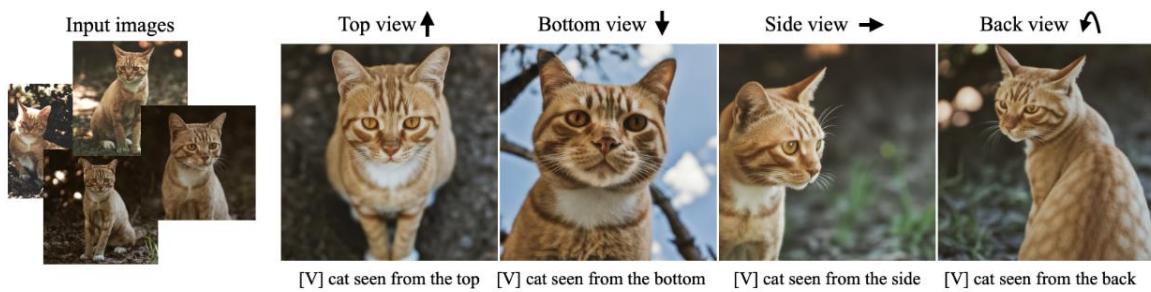
در شکل ۳۷ به خوبی دیده می‌شود که روش‌های مطرح موجود مانند DALL-E و Imagen، در این زمینه بسیار دچار مشکل هستند. تصاویر تولید شده توسط روش ۲ DALL-E و Imagen، صرفا یک تعداد ساعت زرد رنگ تولید کرده اند و طرح ساعت‌های تولید شده هم با یکدیگر متفاوت است و هم همگی با تصویر نمونه تفاوت دارند. اما تصاویر تولید شده توسط روش DreamBooth همان‌طور که انتظار می‌رود دقیقا ساعت موجود در تصاویر نمونه را در فضایی دیگر تولید کرده اند. برای مدل‌های ۲ DALL-E و Imagen، ویژگی‌های ظاهری ساعت به صورت متنی و با جزئیات بالا به مدل داده شده اند(مثلا: ساعت زرد دایره‌ای شکل با صفحه سفید و عقربه ...).



شکل ۳۸ - قابلیت تغییر ویژگی‌های یک تصویر در مدل DreamBooth

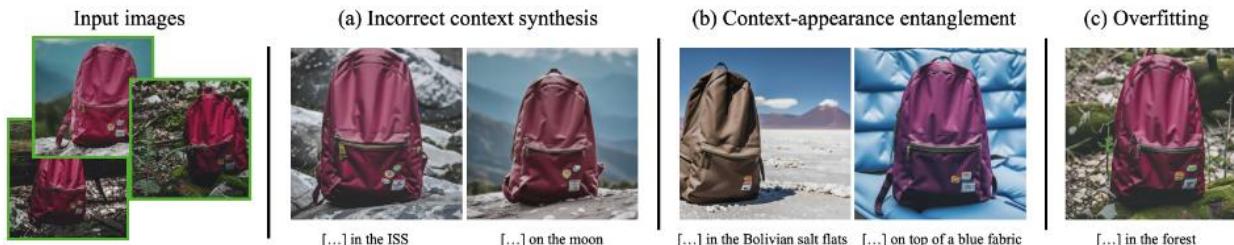


شکل ۳۹ - قابلیت تغییر ویژگی‌های یک تصویر در مدل DreamBooth



شکل ۴۰ - تغییر نمای تصویر و ایجاد تصاویری از نماهای بدیع

اشکال ۳۸ و ۳۹ و ۴۰، ویژگی‌های متعددی که در اثر ثابت بودن سوژه در تصاویر تولیدی به وجود می‌آید را نشان می‌دهد. به نوعی با ثابت نگاه داشتن سوژه در تصاویر، با تغییر متن ورودی می‌توان تصویر را ویرایش کرد که یکی از کاربردهای جالب و پراهمیت این مدل است.



شکل ۴۱ - برخی مشکلات و محدودیت‌های مدل DreamBooth

یکی از مشکلاتی که در این مقاله برای این روش اشاره شده این است که گاهی اوقات، تصاویر ورودی، بر اساس جملات ورودی تغییر می‌کنند. به عنوان مثال در بخش ب شکل ۴۱، رنگ کیف بر اساس اینکه در چه مکانی قرار داشته تغییر کرده است یا در بخش ج وقته از مدل خواسته شده کیف را در جنگل قرار دهد، چون از بین تصاویر نمونه تعدادی تصویر کیف در جنگل وجود داشته، تصاویر تولیدشده بسیار به تصاویر نمونه شبیه شده اند. یکی دیگر از مشکلات طبیعی این مدل زمانی است که بخشی از توضیحات جمله ورودی را به درستی متوجه نشود و تصویر نامطلوبی ارائه دهد.

۵ بحث و جمع بندی

در این گزارش به معرفی مدل‌های پخشی، بررسی کاربردهای آن، اثر آن بر جامعه و علم امروز، نقاط قوت و ضعف و راههایی برای بهبود آن پرداختیم. دیدیم که چگونه این مدل‌ها در طی ۲ الی ۳ سال اخیر تمامی روش‌های تولید تصویر دیگر را منسوخ کرده‌اند و با داشتن ویژگی‌های منحصر به فردی به بهترین روش‌ها برای تولید تصویر تبدیل شده‌اند. مشاهده کردیم که این مدل‌ها در مسائل مختلف مانند ترمیم تصویر، تغییر تصویر، بهبود تصویر، رنگ‌آمیزی و ... مورد استفاده قرار گرفته‌اند ولی یکی از اصلی ترین و پرکاربردترین موارد استفاده این مدل‌ها تبدیل متن به تصویر است. در این چند سال، بسیاری از مردم تولید تصویر به کمک هوش مصنوعی را یکی از انقلاب‌ها و دستاوردهای مهم و ترسناک هوش مصنوعی دانسته‌اند و بسیاری معتقدند این شروع تکامل هوش مصنوعی و بیکار شدن انسان‌ها است. شرکت‌های بسیار بزرگ در حال سرمایه‌گذاری‌های کلان در زمینه این گونه مدل‌ها هستند و به همین دلیل در آینده‌ای نزدیک باید شاهد پیشرفت چندبرابری این مدل‌ها و زمینه تولید تصویر به طور کلی باشیم.

فهرست مراجع

- [1] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "**Denoising diffusion probabilistic models.**" *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.
- [2] Nichol, Alexander Quinn, and Prafulla Dhariwal. "**Improved denoising diffusion probabilistic models.**" International Conference on Machine Learning. PMLR, 2021.
- [3] Saharia, Chitwan, et al. "**Palette: Image-to-image diffusion models.**" *ACM SIGGRAPH 2022 Conference Proceedings*. 2022.
- [4] Ramesh, Aditya, et al. "**Hierarchical text-conditional image generation with clip latents.**" *arXiv preprint arXiv:2204.06125* (2022).
- [5] Rombach, Robin, et al. "**High-resolution image synthesis with latent diffusion models.**" *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
- [6] Ruiz, Nataniel, et al. "**Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation.**" *arXiv preprint arXiv:2208.12242* (2022).