دانشکده مهندسی کامپیوتر

گزارش مطالعاتی درس شبکه های عصبی

شبکه های مولد تقابلی مبتنی بر سبک

استاد درس

جناب آقای دکتر رضا صفابخش

نگارش

محسن عبادپور

(٤٠٠١٣١٠٨٠)

تیرماه ١٤٠١

چکیده

مسئله تولید داده های غیر واقعی و ساختگی اهداف و مقاصد گوناگونی نظیر حل مشکل کمبود داده در یادگیری ماشین و عمیق، کاهش هزینه جمع آوری داده، افزایش تعمیم پذیری مدل، ایجاد تغییر در داده و... را دنبال می‌کند که با معرفی شبکه های مولد تقابلی گام بزرگی در این عرصه برداشته شده است. شبکه های مولد تقابلی با دریافت برداری تصادفی از فضای نهفته[1] اقدام به تولید داده های ساختگی می‌کند که در آن هیچگونه اهرم و ابزار مستقل از هم برای کنترل جزئیات و سبک[2] داده‌ی تولیدی وجود ندارد و این باعث می‌شود نتوان هیچ دیدگاه و چشم اندازی از داده تولیدی تا مرحله پایانی داشت. محققین شرکت Nvidia با معرفی و ارائه شبکه های مولد تقابلی مبتنی سبک توانسته اند فرآیند تولید داده ها را کنترل کرده و تغییرات معنادار و مدنظر را پیش از تولید نهایی داده اعمال کرده و حتی آن را تا حدی پیش‌بینی نمایند. در این گزارش ضمن بررسی و آنالیز اولین نسخه شبکه های مولد تقابلی مبتنی بر سبک، استفاده از آن برای سه مسئله‌ی ترجمه تصویر به تصویر، ویرایش معنی‌دار و قطعه‌بندی شی در تصاویر مورد مطالعه قرار گرفته و در نهایت نسخه دوم شبکه های مولد تقابلی مبتنی بر سبک معرفی شده و فضای سبک آن تحلیل و گزارش خواهد شد.

کلمات کلیدی: انتقال سبک[3]، دستکاری معنادار[4]، عادی سازی تطبیقی نمونه[5]، شبکه هم‌نهاد[6]، مدل های مولد

[1] Latent space

[2] Style

[3] Style transfer

[4] Meaningful manipulation

[5] Adaptive instance normalization

[6] Synthesis network

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل ۱ – مقدمه

تولید داده های غیر واقعی و ساختگی به مسئله مهمی در دهه اخیر تبدیل شده است بطوری که با وجود نیازمندی به منابع پردازشی قدرتمند برای آموزش مدل های مولد، همچنان دست آورد های ناشی از آن ارزش صرف منابع را دارد. در این فصل به اهمیت مدل های مولد به همراه تاریخچه مختصری از آن پرداخته شده و شبکه مولد تقابلی معرفی و چالش موجود برای آن ترسیم می‌شود.

۱ – ۱- نیازمندی به مدل های مولد

نیازمندی روز افزون به داده های بیشتر برای مقاصد مختلف امری اجتناب ناپذیر است و بایستی با کمترین هزینه پردازشی، مالی و با بیشترین کیفیت ممکن حجم داده ها را افزایش داده یا در آن تغییر ایجاد نمود. تولید داده ضمن جلوگیری از کمبود آن برای یادگیری، باعث صرفه جویی در هزینه های جمع آوری و برچسب گذاری داده های واقعی شده و افزایش صحت پیش‌بینی برای مدل های یادگیری هدف(دسته بندی و...) را در پی دارد.

همچنین تولید داده باعث اجتناب از بیش‌برازش شده و به ما این امکان را می دهد که مدل‌های متنوع و تعمیم‌پذیرتری را ارائه نموده و مشکل نامتعادل[۷] بودن کلاس های مختلف در مسئله‌ی دسته بندی را حل کرد. بنابر توضیحات ارائه شده می‌توان انتظار داشت که هر چقدر فرآیند تولید داده توسط مدل های مولد دقیق، معتبر و قابل اتقا باشد، به همان میزان داده های تولیدی قدرت بالایی در اهداف مذکور خواهند داشت لذا نقش مهم مدل های مولد و اهمیت دقت آن آشکار می‌شود.

۱ – ۲- اولین مدل های مولد

در علم شناسایی کلاسیک الگو، مسئله‌ی تخمین توزیع و پارامتر موضوع مهمی تلقی شده و نقش پررنگی در انواع مباحث مربوطه نظیر دسته‌بندی و... دارد. اگر بتوان به ازای داده های کلاسی از مجموعه داده‌ی مد نظر، توزیع و پارامتر های مربوط به آن را محاسبه یا تخمین زد، در آن صورت می‌توان به ازای مقادیر دلخواه دیگر و بر اساس توزیع حاصل، داده های جدید از همان کلاس را تولید کرد لذا اولین مدل های مولد را می‌توان مدل های آماری مبتنی بر تخمین توزیع داده‌ها دانست؛ از تخمین‌گر های آماری برای این هدف می‌توان به تخمین‌گر پارامتر بیزی[۸] اشاره نمود که از آن برای تخمین پارامتر های توزیع مجموعه داده‌های هدف استفاده می‌شود.

[7] Imbalance

[8] Bayesian Parameter Estimation

۱ - ۳- شبکه مولد تقابلی

با پیدایش شبکه های عصبی و رشد چشمگیر آن در دو دهه اخیر، شاهد پیشرفت های مهم و متعددی در مباحث یادگیری بوده‌ایم. یادگیری برای تولید داده‌های جدید با معرفی شبکه های مولد تقابلی [۱] وارد نسل جدیدی شده و نتایج بی‌نظیر آن محققین این عرصه را به پژوهش در این زمینه تشویق کرده است. شبکه مولد تقابلی از دو شبکه عصبی مجزا از هم با عناوین "تمایزگر"[۹] و "مولد"[۱۰] تشکیل شده است؛ شبکه تمایزگر یک دسته بند دودوئی بوده که هدف آن تشخیص داده های تولیدی از داده های واقعی می‌باشد و شبکه مولد نیز یک شبکه برای تولید داده می‌باشد که با دریافت برداری تصادفی از فضای نهفته، آن را به داده‌ی تولیدی مد نظر(تصویر و...) تبدیل می‌کند که ویژگی های ورود به شبکه تمایزگر را داراست که در بحث تولید تصویر این ویژگی ها شامل ابعاد و تعداد کانال ها می باشد.

آموزش این دو شبکه بصورت توامان انجام می‌پذیرد که در آن شبکه مولد سعی می‌کند داده هایی تولید کند که شبیه داده های واقعی بوده و تمایزگر نتواند آن را شناسایی کند و شبکه تمایزگر نیز تلاش می‌کند که به بهترین نحو داده های تولیدی از داده های واقعی را تفکیک نماید. از جایی که تابع هزینه‌ی دو شبکه‌ی مذکور در مقابل هم می‌باشد، در فاز آموزش بین شبکه های مولد و تمایزگر رقابتی نظیر بازی کمینه/بیشینه[۱۱] رخ می‌دهد و شبکه مولد از این رقابت یاد می‌گیرد چگونه با دریافت برداری تصادفی، داده‌ای تولید نماید که شبیه داده های واقعی بوده و در توزیعی نظیر آن قرار داشته باشد. پس از اتمام آموزش این دو شبکه بصورت همزمان، می توان هر یک را بصورت مستقل از هم به کار برد و از مولد برای اهداف تولید داده ها استفاده نمود.

چالشی که در شبکه های مولد تقابلی وجود دارد عدم امکان کنترل در داده های تولیدی و ویژگی های موجود در آن می باشد. اگر مسئله تولید تصاویر چهره‌ی انسانی را در نظر بگیریم، در شبکه های مولد تقابلی نمی توان هیچگونه کنترل یا اثری در جزئیات و خصوصیات چهره‌ی تولیدی قائل بود بطوری که تا زمان تولید نهایی تصویر نمی‌توان گفت که چهره تولیدی آیا مرد است یا زن، سیاه پوست است یا سفید پوست، چشم مشکی است یا چشم رنگی، عینکی است یا خیر، کلاه بر سر دارد یا خیر، در محیط تاریک و کم نور است یا روشن و ... . در پاسخ به این چالش شبکه مولد تقابلی مبتنی بر سبک [۲] معرفی شده است که در فصل بعد معرفی و مورد بررسی قرار خواهد گرفت.

[9] Discriminator

[10] Generator

[11] Min/Max

فصل ۲ – شبکه مولد تقابلی مبتنی بر سبک

کنترل بر ویژگی و خصوصیات داده های تولیدی از اهمیت بسزایی برخوردار است چرا که این امکان را به ما می‌دهد که بتوانیم پیش از تولید داده، ویژگی های سطح بالای آن را تا حدی پیش‌بینی نموده و بر المان های آن کنترل داشته باشیم و این امری است که جای خالی آن در شبکه های مولد تقابلی احساس می شود چرا که فضای نهفته همانند جعبه‌ی سیاهی بوده که ما از آن اطلاعی نداریم. شبکه مولد تقابلی مبتنی بر سبک(به اختصار StyleGAN) در پاسخ به خواسته مذکور معرفی شده است که در این فصل مورد بررسی قرار خواهد گرفت. از جایی که اساس و توضیحات ارائه شده در مقاله مربوطه[۲] بر اساس داده های تصویری و مبتنی بر چهره می باشد، گزارش موجود نیز با فرض اینکه داده تولیدی تصویر می باشد ارائه می‌گردد.

۲ – ۱- تمایزگر و تابع هزینه

در شبکه مولد تقابلی مبتنی بر سبک هدف کنترل ویژگی های سطح بالای داده تولیدی بوده و تمرکز بر تغییر شبکه مولد قرار گرفته است لذا شبکه تمایزگر و ساختار پایه‌ی تابع هزینه در StyleGAN همانند شبکه مولد تقابلی بوده و تغییری در آن به وجود نیامده است و تمایزگر بیش از پیش همانند یک دسته‌بند دودوئی عمل می کند.

۲ – ۲- استقلال، تفسیر پذیری و اهمیت فضای نهفته

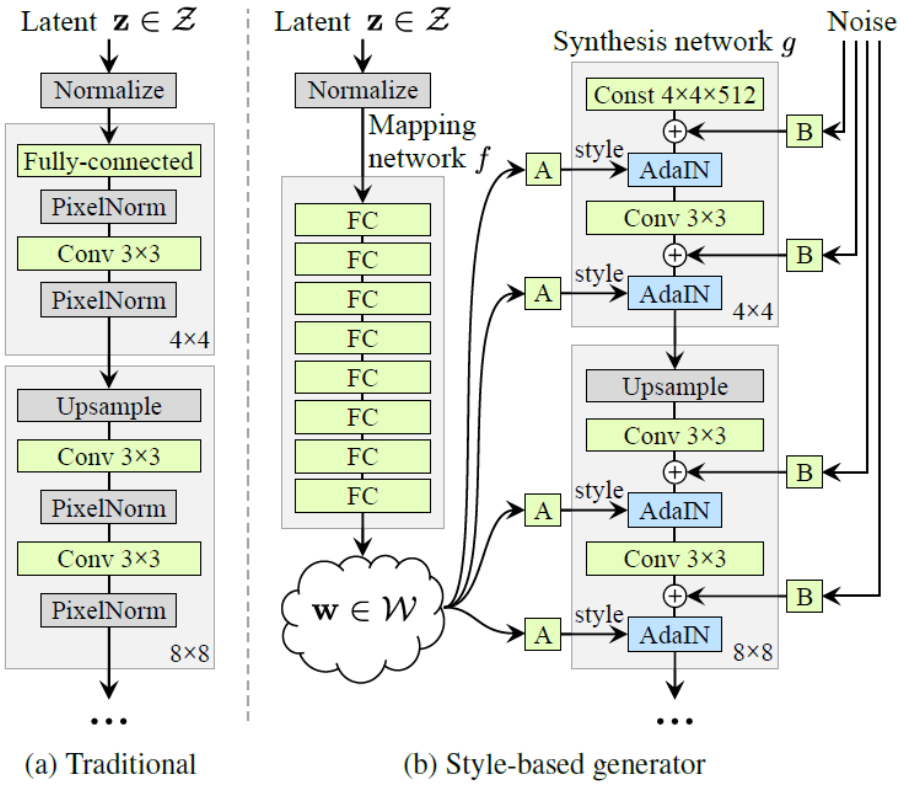
میتوان گفت تاثیر فضای نهفته در تولید داده های ساختگی انکار ناپذیر است بطوری که متغیر های آن، ویژگی های سطح بالای داده‌ی تولیدی را تعیین می‌کند؛ از این رو برای محققین مهم است که بتوانند فضای نهفته و متغیر های آن را تفسیر کنند تا داده‌ی تولیدی با ویژگی های مربوطه مطابق میل حاصل شود. از طرفی تفسیر پذیری فضای نهفته باعث می‌شود که بتوان شبکه های مولد مختلف را در جایگاه مقایسه با یکدیگر قرار داد لذا هر چه فضای نهفته شناخته شده‌تر باشد، بهتر می‌باشد. همچنین مهم است که تاثیر متغیر های فضای نهفته در داده تولیدی تا حد امکان از هم مستقل و مشخص باشد چرا که بتوان با تغییر یک متغیر، تغییر مطلوب و مد نظر در داده‌ی تولیدی را ایجاد نمود؛ در StyleGAN نیز با تعبیه شبکه نگاشت[۱۲] در معماری مولد و همچنین ترکیب منظم‌ساز[۱۳] این هدف دنبال می شود.

همانطور که در شکل (۱) مشخص شده است، شبکه نگاشت یک شبکه عصبی جلو-روی هشت لایه می‌باشد که با دریافت بردار تصادفی از فضای نهفته z، آن را به برداری در فضای نهفته جدید w نگاشت می کند. این نگاشت باعث می شود فضای نهفته جدید ویژگی مهم مذکور را دارا باشد(استقلال نسبی متغیر ها). ترکیب منظم‌ساز نیز به منظور جلوگیری از

[12] Mapping network

[13] Mixing regularization

وابستگی متغیر های فضای نهفته جدید اعمال شده است که در آن از دو یا چند بردار نهفته متفاوت(W_1، W_2 و... متعلق به Z_1، Z_2 و...) برای تولید تصویر استفاده می شود.



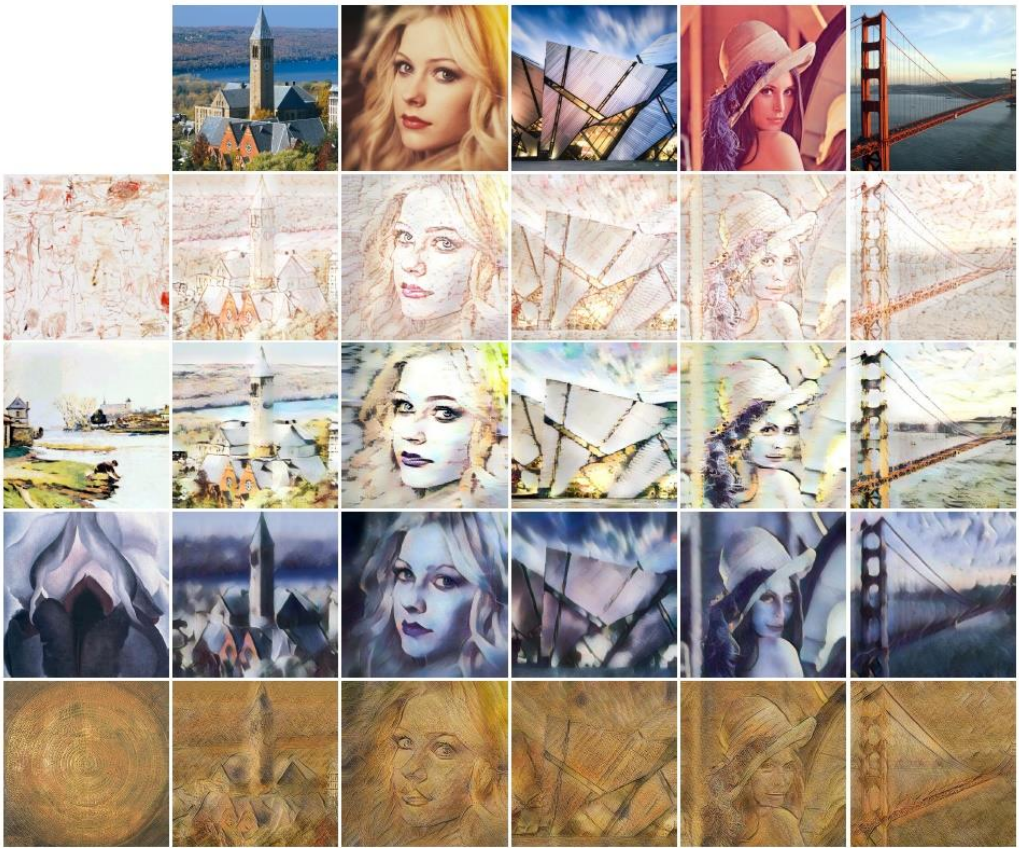
شکل ۱: مقایسه معماری شبکه مولد و ساختار آن در شبکه های مولد تقابلی مبتنی بر سبک و مولد تقابلی [۲]

۲ – ۳- انتقال سبک و عادی سازی تطبیقی نمونه

انتقال سبک و ساختار از تصویری به تصویر دیگر و ترکیب سبک های مختلف با هم می تواند زمینه‌ساز تولید تصاویر ساختگی جدید با حفظ آن ساختار ها باشد. عادی سازی(نرمال سازی) تطبیقی نمونه[14] (به اختصار AdaIN) با هدف انتقال سبک در بخشی از پژوهش هوآنگ و همکاران [۳] مورد استفاده قرار گرفته و نشان داده اند که اگر تصویر ورودی بر اساس مقادیر ویژگی های خود نرمال شده و سپس توسط میانگین و واریانس تصویر مرجع مقیاس شوند، سبک و ساختار تصویر مرجع به تصویر ورودی انتقال پیدا می‌کند. رابطه (۱) نشان دهنده AdaIN بوده که در آن x مربوط به تصویر ورودی و y مربوط به تصویر مرجع می باشد. همچنین شکل (۲) نمونه هایی از اعمال انتقال سبک بین دو تصویر متفاوت را نشان می‌دهد.

$$\mathrm{AdaIN}(x, y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right) + \mu(y) \qquad (۱)\ [۳]$$

[14] Adaptive instance normalization



شکل ۲: نمونه هایی از انتقال سبک بین تصاویر [۳]

۲ – ۴- شبکه مولد و معماری آن در شبکه مولد تقابلی مبتنی بر سبک

در StyleGAN پنج تغییر اساسی نسبت به حالت پایه(A) پیشنهاد شده است که هر کدام باعث بهبود کیفیت تصاویر ضمن کنترل سبک تصاویر تولیدی گردیده است که در ادامه مورد بررسی قرار می گیرد؛ اولین تغییر(B) استفاده از درون یابی دو خطی[۱۵] برای فرا/فرو نمونه برداری[۱۶] می باشد که در نسخه های قبلی از سایر روش ها نظیر لایه پیچشی معکوس[۱۷] استفاده شده است. این تغییر باعث افزایش سرعت محاسبات و کاهش پیچیدگی برای یادگیری پارامتر ها شده و ارتباط محلی بین پیکسل های همسایه را حفظ می‌کند.

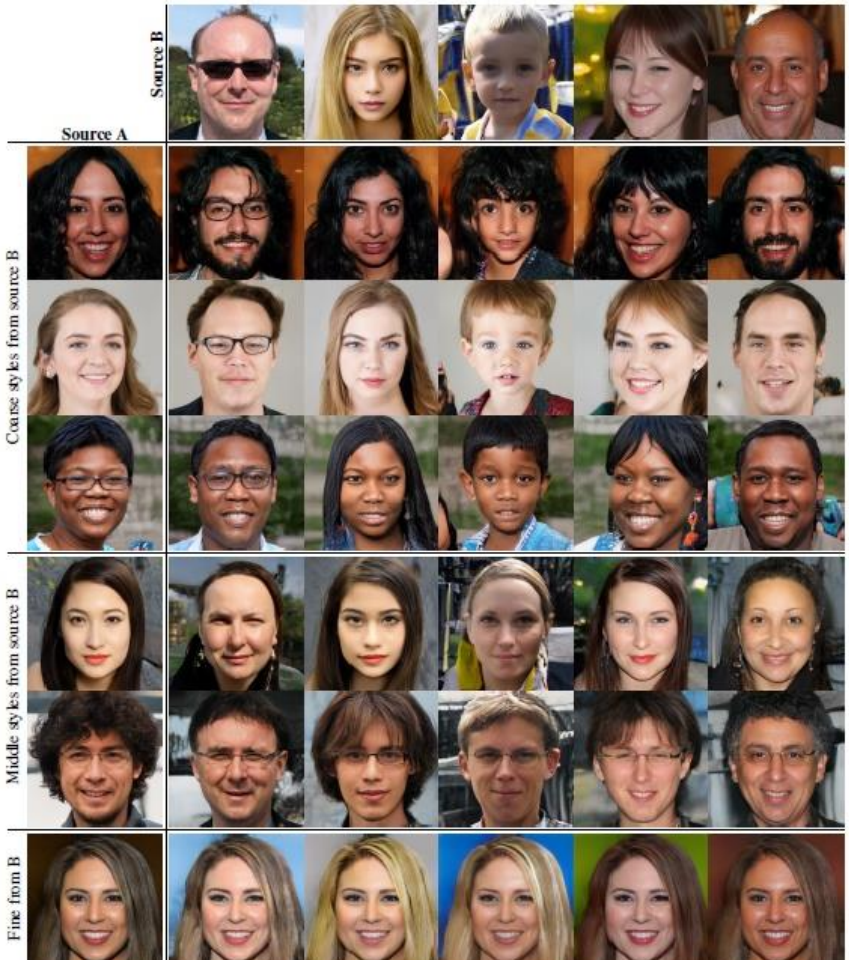
دومین تغییر(C) اضافه کردن شبکه نگاشت و انتقال سبک به تصویر تولیدی می باشد که در هر مرحله از فرا نمونه برداری یک سبک به تصویر موجود انتقال داده می شود که مقادیر سبک انتقالی از w می‌باشد؛ سبک های انتقالی در مراحل

[15] Bilinear

[16] up/downsampling

[17] Transposed Convolution

ابتدایی فرا نمونه برداری، ویژگی های پایه نظیر فرم ایستادگی صورت، حالت کلی مو، شکل چهره، و... را انتقال می دهند و هر چقدر انتقال سبک به گام های پایانی نزدیک باشد، ویژگی های کلی انتقال داده می‌شود. شکل (۳) تاثیر انتقال سبک در هر مرحله از فرا نمونه برداری را نشان می‌دهد.



شکل ۳: تاثیر انتقال سبک در هر یک از مراحل تولید تصویر چهره [۲]

سومین تغییر(D) حذف بردار ورودی به شبکه مولد می‌باشد؛ پیشنهاد ارائه شده در مقاله [۲] این می‌باشد که به جای تولید تصادفی یک بردار ورودی برای تولید تصویر ساختگی جدید، بردار ورودی ثابت مانده و صرفا سبک های انتقالی به تصویر موجود تغییر یابد که این پیشنهاد باعث می‌شود قسمتی از حجم پردازش‌ها و پارامتر های یادگیری در مرحله محاسبه بردار ورودی کاهش یافته و بررسی تاثیر متغیر های فضای نهفته w در سبک تصاویر تولیدی راحت تر شود. چهارمین تغییر (E) اضافه کردن نویز به تصویر قبل از هر انتقال سبک می‌باشد که این باعث می شود تصاویر واقعی‌تر حاصل شده و از حالت انیمیشنی دور شود و همچنین سبک های انتقالی متوالی به هم نابسته حاصل شوند. پنجمین تغییر(F) استفاده از ترکیب دو یا چند بردار نهفته w برای تولید یک تصویر ساختگی می‌باشد که این مورد نیز به جهت استقلال سبک های انتقالی پیشنهاد گردیده است.

۲ – ۵- آموزش شبکه StyleGAN و نتایج تجربی

آموزش شبکه مولد تقابلی مبتنی بر سبک توسط چند مجموعه داده متفاوت انجام پذیرفته است که دو مورد از اهم آنها مجموعه داده‌ی CelebA-HQ [۴] و Flickr-Faces-HQ [۲] می باشد. شاخص ارزیابی نیز Frechet Inception Distance [۷] (به اختصار FID) می باشد که یک معیار مبتنی بر فاصله برای ارزیابی فاصله توزیع تصاویر تولید شده توسط شبکه مولد با توزیع تصاویر آموزشی می باشد و کمینه بودن این فاصله حاکی از قدرت بالای شبکه مولد است. به ازای هر کدام از تغییرات پیشنهادی برای شبکه مولد در [۲]، آموزش شبکه انجام شده و میزان تاثیر هر یک با فاصله FID گزارش شده است که جدول (۱) نتایج تجربی حاصل از آن را نشان می‌دهد.

جدول ۱: نتایج تجربی حاصل از آموزش شبکه مولد تقابلی مبتنی بر سبک [۲] در دو مجموعه داده متفاوت

	Method	CelebA-HQ	FFHQ
A	Baseline Progressive GAN [30]	7.79	8.04
B	+ Tuning (incl. bilinear up/down)	6.11	5.25
C	+ Add mapping and styles	5.34	4.85
D	+ Remove traditional input	5.07	4.88
E	+ Add noise inputs	**5.06**	4.42
F	+ Mixing regularization	5.17	**4.40**

نتایج آموزش نشان می‌دهد هر یک از تغییرات پیشنهادی می تواند در مجموعه داده های مختلف عملکرد متفاوتی داشته باشد اما آنچه مسلم برآیند آن یک حرکت رو به رشد بوده و استفاده‌ی آن دست آورد های متعددی را به همراه دارد که اهم آن کنترل بر سبک داده‌ی تولیدی و ایجاد تغییر در آن می باشد.

فصل ۳ – ترجمه تصویر به تصویر مبتنی بر سبک

در فصل قبل ضمن معرفی شبکه های مولد تقابلی مبتنی بر سبک ملاحظه کردیم که می‌توان با داشتن یک تصویر ثابت ورودی و با انتقال سبک های مختلف در مراحل متفاوت فرا نمونه برداری تصاویر گوناگونی تولید نمود. حال می‌توان گفت که با داشتن سبک های مقتضی برای یک تصویر مشخص، امکان آن وجود دارد که تصویر را بازسازی نموده یا در آن تغییراتی حتی معنادار به وجود آورد. در این فصل نتیجه‌ی یک پژوهش مورد بررسی قرار خواهد گرفت[۵] که هدف آن پاسخ گویی به مسائل ترجمه تصویر به تصویر نظیر بازسازی تصویر تخریب شده، افزایش وضوح تصویر و... با استفاده از StyleGAN می باشد.

۳ - ۱- ایجاد تغییرات معنادار

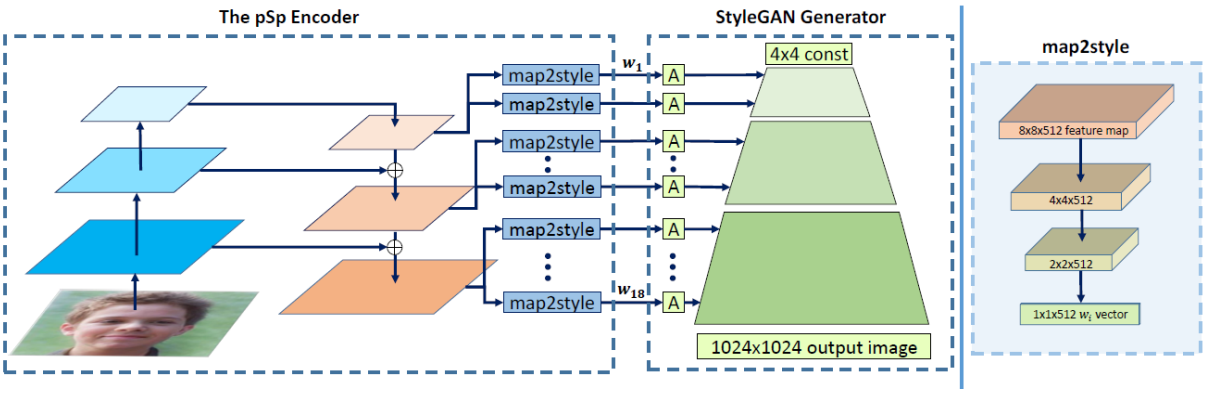
در StyleGAN اثبات شد که تصاویر گوناگون حاصل از تصاویر یکسانی هستند که سبک های مختلف به آن انتقال داده شده است؛ حال اگر تصویر موجود و مد نظر خود را به صورت معکوس در شبکه مولد پیش بریم، به بردار نهفته معادل با w برای تصویر مفروض دست خواهیم یافت و اکنون در صورتی که تغییراتی در w (سبک های تصویر) ایجاد کرده و دوباره تصویر را بازسازی نماییم، تغییر اعمالی در تصویر نهایی نیز ظاهر خواهد شد و از طرفی در [۲] مشاهده کرده‌ایم که در هر مرحله سبک های انتقالی در چه سطحی به عکس اثر میگذارد لذا آگاهی لازم از اعمال تغییر در سبک های هدف وجود داشته و مسیر تقریبا همواری از ایجاد تغییر در تصاویر ترسیم می شود؛ اما آنچه مسلم است تبدیل معکوس یاد شده و بازسازی تصویر بصورت کامل امری چالشی بوده و از دقت کاملی برخوردار نمی‌باشد لذا یک مشکل همواره موجود چگونگی تبدیل یک تصویر به فضای نهفته w می باشد که در مقاله مد نظر [۵] به آن پرداخت شده و یک معماری جدید کدگذاری معرفی گردیده است؛ معماری معرفی شده در بخش های (۳-۲) و (۳-۳) مورد بررسی قرار خواهد گرفت.

۳ - ۲- معماری و ساختار شبکه پیکسل-سبک-پیکسل

معماری های مختلفی در حیطه ترجمه تصویر به تصویر معرفی و ارائه شده است که اکثریت آن امکان توسعه به مسائل مختلف را نداشته یا نیازمند آموزش مجدد مولد و بهینه سازی آن می باشد؛ روش پیشنهادی در [۵] شامل استفاده از شبکه مولد StyleGAN از پیش آموزش دیده شده می باشد؛ همچنین روند کاری آن بصورت پردازش های پیکسل-سبک-پیکسل[18] (به اختصار pSp) است. در pSp ابتدا تصویر ورودی به فضای نهفته +w نگاشت می شود(پیکسل به سبک) و سپس تغییرات مد نظر در فضای نهفته اعمال شده و مجددا توسط سبک موجود تصویر بازسازی می‌شود(سبک به پیکسل). ضمنا در پژوهش مذکور مقصود از فضای نهفته +w ترکیب ۱۸ بردار متفاوت ۵۱۲ بُعدی از فضای نهفته w شبکه StyleGAN برای تخمین سبک های انتقالی می باشد.(به ازای ۱۸ لایه‌ی فرا نمونه برداری).

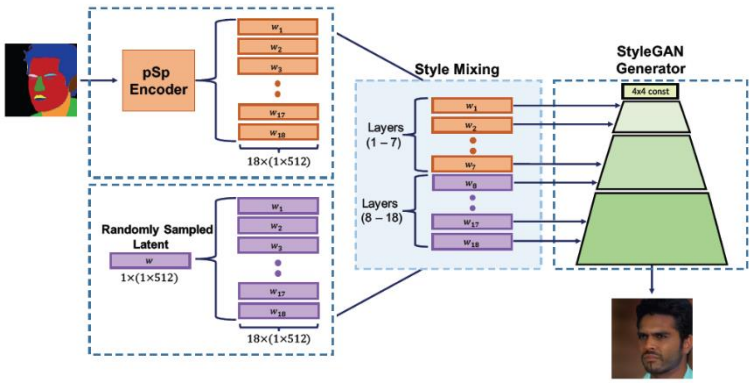
[18] pixel2Style2pixel

شکل (۴) معماری ارائه شده در پژوهش [۵] را نشان می دهد که ابتدا با استفاده از هرم ویژگی[19] [۶] فرآیند استخراج ویژگی از تصویر ورودی انجام شده و سپس توسط یک شبکه کانولوشنی بردار های معادل در فضای نهفته w+ حاصل می شود که نگاشتی از تصویر به فضای مذکور می باشد. شبکه مولد نیز با دریافت بردار های حاصل اقدام به بازسازی تصویر می‌نماید؛ در این گام میتوان تغییرات مد نظر را در بردار ها را اعمال نمود تا تصویر حاصل متحول تغییرات شود.



شکل ٤: معماری ترجمه تصویر به تصویر[۵]

شکل (۵) یک نمونه از چگونگی استفاده از pSp برای پاسخ به مسائل ترجمه تصویر به تصویر می‌باشد؛ تصویر داده شده یک تصویر قطعه‌بندی[20] شده می باشد و هدف تولید تصویر واقعی بر اساس قطعه بندی انجام شده می باشد. برای این کار تصویر ورودی به فضای w+ نگاشت شده و از آن صرفا برای انتقال سبک در لایه های ابتدایی استفاده می شود چرا که فرم و حالت تصویر در لایه های ابتدایی تعیین می‌شود(قسمت قطعه بندی شده). سپس به لایه های بعدی، بردار هایی تصادفی از فضای w+ برای انتقال سبک تزریق می شود تا یک تصویر ساختگی تولید شود؛ بدین صورت ضمن حفظ نواحی قطعه‌بندی شده، جزئیات جدید به آن اضافه می شود.



شکل ۲: نمونه ای از چگونگی اعمال تغییر در تصویر و ترجمه تصویر به تصویر [۵]

[19] Feature pyramid

[20] Segmented image

۳ – ۳– آموزش شبکه پیکسل-سبک-پیکسل و نتایج تجربی

آموزش شبکه pSp توسط تابع هزینه رابطه (۲) انجام و وزن های شبکه کدگذار[۲۱] به هنگام در آورده می‌شود. این تابع هزینه از ترکیب وزن‌دار چهار جمله تشکیل شده است که بترتیب مربوط به خطای مبتنی بر پیکسل، شاخص Learned Perceptual Image Patch Similarity [۱۸] (به اختصار LPIPS)، شاخص یکتا بودن (مبتنی بر شبکه Arcface [۱۹]) و جمله منظم‌ساز نسبت به فضای نهفته w می باشد.

$$\mathcal{L}(\mathbf{x}) = \lambda_1\mathcal{L}_2(\mathbf{x}) + \lambda_2\mathcal{L}_{\text{LPIPS}}(\mathbf{x}) + \lambda_3\mathcal{L}_{\text{ID}}(\mathbf{x}) + \lambda_4\mathcal{L}_{\text{reg}}(\mathbf{x}) \quad (۲)\ [۵]$$

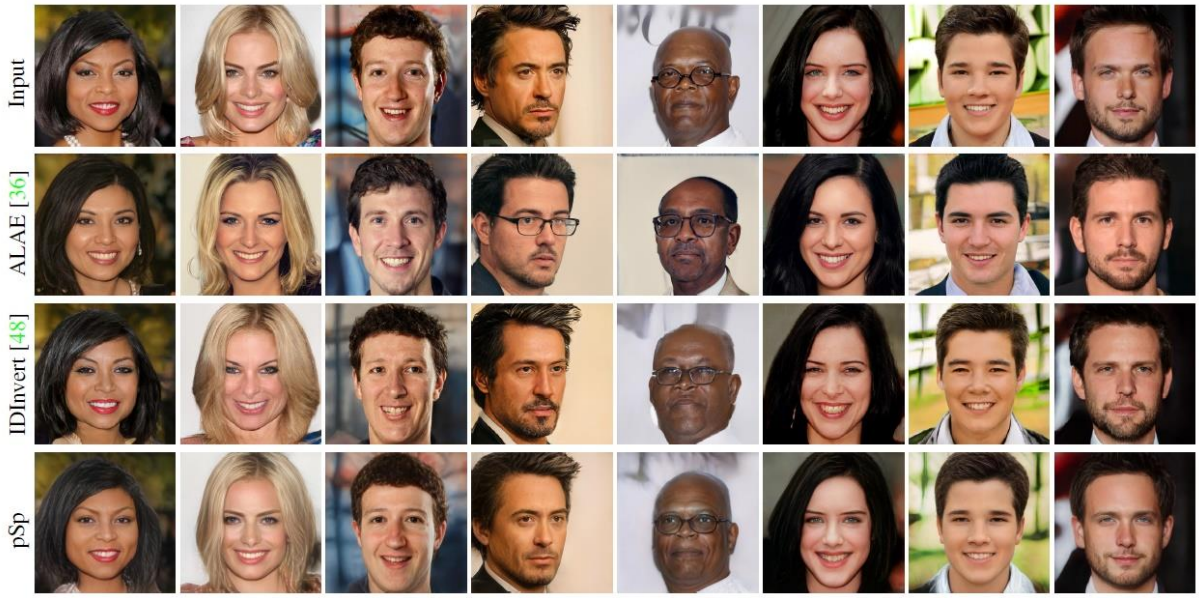
برای مقایسه معماری پیشنهاد شده در پژوهش [۵] نتایج حاصل از بازسازی تصویر ورودی با سه روش دیگر مقایسه و در جدول (۲) آورده شده است؛ همچنین نمونه تصاویر مقایسه شده نیز در شکل (۶) قابل ملاحظه است. از جدول (۲) می توان دریافت که عملکرد معماری pSp به جهت شاخص های کیفی در رقابت تنگانگی با معماری [۹] قرار دارد اما برتری ماورایی pSp در زمان اجرایی باعث آشکار شدن قدرت آن می‌شود.

جدول ۲: مقایسه عملکرد معماری pSp با سایر متد های موجود برای ترجمه تصویر به تصویر [۵]

Method	↑ Similarity	↓ LPIPS	↓ MSE	↓ Runtime
Karras *et al.* [22]	0.77	0.11	0.02	182.1
ALAE [36]	0.06	0.32	0.15	0.207
IDInvert [48]	0.18	0.22	0.06	0.032
$\mathcal{W}$ Encoder	0.35	0.23	0.06	0.064
Naive $\mathcal{W}+$	0.49	0.19	0.04	0.064
pSp w/o ID	0.19	0.17	0.03	0.105
pSp	0.56	0.17	0.03	0.105

تصاویر حاضر در شکل (۶) نیز نشان می‌دهد که عکس های بازسازی شده توسط pSp جزئیات بیشتری را حفظ و اصالت ویژگی های تصویر اصلی را حمل کرده است.

[21] Encoder



شکل ۳: نمونه تصاویر بازسازی شده توسط pSp و سایر معماری های موجود [۵]

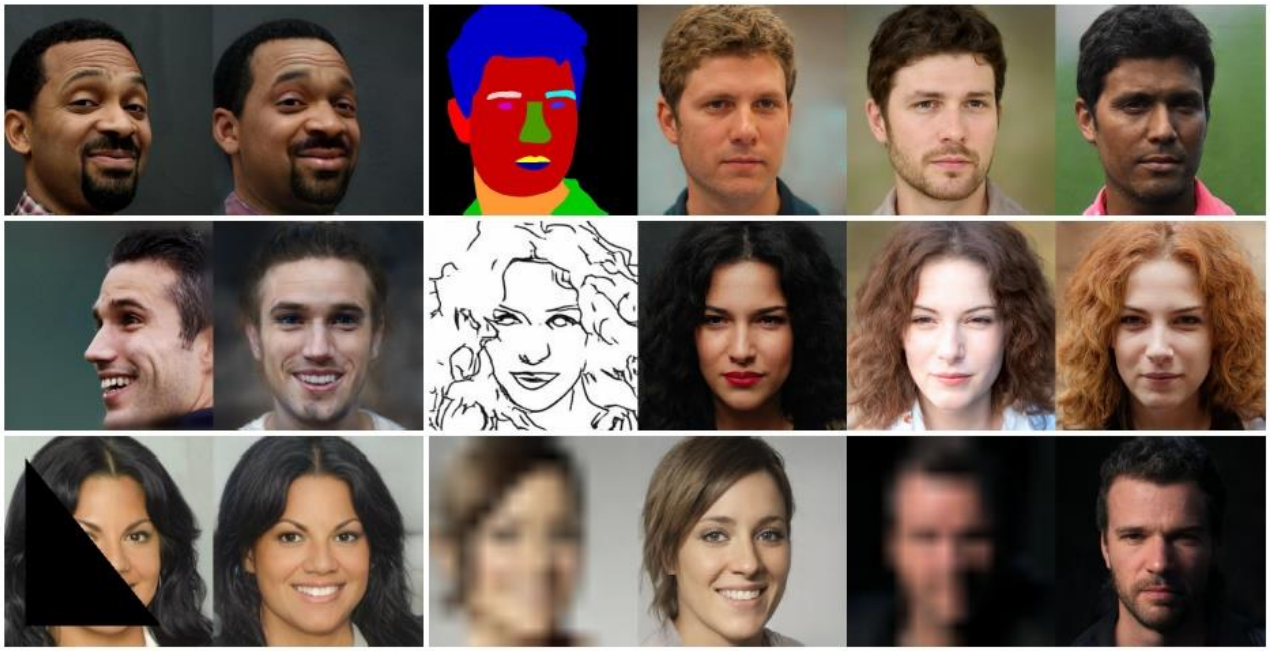
۳ - ۴- برتری و گسترش معماری پیکسل-سبک-پیکسل در سایر مسائل

همانطور که در بخش قبل ملاحظه شد، معماری pSp ضمن سرعت پاسخگویی بالا و کیفیت بازسازی مطلوب امکان تغییر در تصویر و یا ترجمه تصویر به تصویر را فراهم می‌کند و از جایی که شبکه مولد جداگانه‌ای آموزش داده نشده و از شبکه مولد از پیش آموزش دیده شده استفاده می‌کند، هزینه محاسباتی pSp بالا نبوده و توجیه استفاده و امکان بسط در سایر مسائل را پیدا می‌کند؛ همچنین توان بالا و دقیق pSp در نگاشت تصویر ورودی به فضای نهفته w امکان ایجاد دستکاری یا تغییر معنادار هر چه بهتر در آن فضا را فراهم می‌کند. مسائل متنوع مطرح در [۵] علاوه بر بازسازی تصویر از فضای نهفته، شامل موارد متعددی نظیر افزایش وضوح[۲۲]، رنگ آمیزی[۲۳] تصویر، ترسیم چهره از رو به رو[۲۴] و تبدیل نقاشی به تصویر واقعی و... می باشد که در اشکال (۷) نمونه هایی از آن آورده شده است. در مسئله رنگ آمیزی نقاشی و تبدیل آن به تصویر واقعی قابل مشاهده است که به ازای بردار های تصادفی مختلف، تصاویر تولیدی می‌تواند از هم فاصله بگیرد بطوری که در تولید چهره خانم در ردیف میانی شکل (۷) این نکته قابل ملاحظه است.

[22] super-resolution

[23] Inpainting

[24] Facial frontalization



شکل ٤ : نمونه هایی از کاربرد pSp در مسائل ترجمه تصویر به تصویر[٥]

فصل ۴ – ویرایش تصاویر مبتنی بر سبک

در فصل قبل ملاحظه کردیم که اگر بتوان تصویر ورودی و مد نظر را به فضای نهفته W یا W^+ نگاشت کرد بطوری که بازسازی آن با خطای حداقلی انجام گیرد، می‌توان ویرایش و دستکاری های معنادار در فضای مذکور ایجاد کرده و نتیجه را در تصویر خروجی دریافت نمود. در حال حاضر چالش نگاشت یک مسئله مهم بوده و محققین در حال پژوهش در این حوزه می‌باشند. در این فصل، نتایج پژوهش [۸] مورد بررسی و گزارش قرار خواهد گرفت که تلاش اصلی بر معرفی یک کدگذار برای نگاشت تصاویر به فضای نهفته با هدف ویرایش آن می‌باشد.

۴ – ۱- انواع نگاشت تصویر به فضای نهفته

اگر نگاشت تصویر ورودی به فضای نهفته و معکوس آن در بازسازی تصویر ضعیف انجام شود، انجام ویرایش مد نظر نیز با مشکل مواجه شده و ویرایش های اعمالی می‌تواند نتایج کنترل نشده‌ای را پیش آورد. سه رویکرد برای این هدف به کار گرفته شده است که عبارت است از:

- مسئله بهینه سازی: تولید تصویر توسط بردار های فضای W بصورت مستقیم یک مسئله بهینه سازی در نظر گرفته شده و هدف کمینه کردن خطای بازسازی تصویر بر اساس بردار انتخابی از آن فضا بوده و روشی پر هزینه و زمانبر می‌باشد.
- آموزش کدگذار: در این رویکرد یک کدگذار مجزا آموزش داده می‌شود تا بتواند تصاویر ورودی را به فضای نهفته نگاشت کند بطوری که خطای بازسازی به عنوان تابع هزینه آن در نظر گرفته می‌شود.
- ترکیبی: از دو رویکرد فوق بصورت توامان استفاده می‌شود.

از جایی که در پژوهش [۸] هدف امکان ویرایش تصاویر در فضای نهفته می‌باشد، محققین مقاله تلاش کرده‌اند نگاشت تصویر ورودی مبتنی بر آموزش کدگذار باشد که در بخش(۴-۳) این کدگذار مورد بررسی قرار می‌گیرد.

۴ – ۲- دو راهی فضای نهفته W و W^+

در پژوهش [۵] دیدیم که برای بازسازی بهتر تصویر ورودی و تخمین راحت‌تر سبک های تصویر از فضای W^+ بهره گرفته شده و نتایج آن نیز قابل قبول بود؛ اما چالشی که در استفاده از فضای نهفته W^+ وجود دارد، محدودیت در انتخاب سبک انتقالی و ایجاد ویرایش مد نظر در تصویر ورودی می‌باشد چرا که به جای یک بردار، چندین بردار بایستی تغییر یابند و طبیعتا تعیین تغییر در چند بردار و ارتباط آن تغییر با ویرایش سطح بالای تصویر، پیچیده‌تر از تعیین تغییر در یک بردار می باشد و این باعث می‌شود بین ویرایش پذیری تصویر با بازسازی آن یک دو راهی به وجود آید که همان نیز عبارت است از اینکه نگاشت به W انجام شود یا W^+. از جایی که در پژوهش [۸] هدف ویرایش تصاویر در فضای سبکی می‌باشد، کدگذار پیشنهادی (که

در بخش ۴-۳ مورد بررسی قرار می‌گیرد) سعی در نگاشت تصویر به فضای نهفته w دارد. برای نشان دادن ادعای فوق که امکان ویرایش تصویر و معنادار بودن آن در فضای w بهتر می باشد، یک آزمایش در [۸] انجام شده است که در شکل (۸) قابل مشاهده است.

شکل ۵ : نتایج اعمال ویرایش در تصاویر در مقایسه بین فضای w و +w [۸]

در شکل (۸) اگر سطر میانی که مربوط به اعمال ویرایش در فضای سبکی +w می باشد را در نظر بگیریم، می‌بینیم با وجود اینکه نتیجه ویرایش خودرو به تصویر اصلی خود نزدیک‌تر می‌باشد(بازسازی بهتر) اما ویرایش اعمالی باعث شده است که تصویر خودروی حاصل غیرعادی شود؛ مثلا کشیدگی کاپوت یک مورد غیر عادی می‌باشد. حال اگر سطر پایین که مربوط به اعمال ویرایش در فضای سبک w می‌باشد را در نظر بگیریم، ملاحظه می‌کنیم که تصاویر حاصل با وجود اینکه به خودروی ورودی شبیه نمی‌باشد اما تصویر حاصل پس از اعمال ویرایش کاملا عادی و معنادار است لذا انتخاب فضای w و سعی در نگاشت تصویر ورودی به آن جهت اعمال ویرایش، انتخابی درست می باشد.

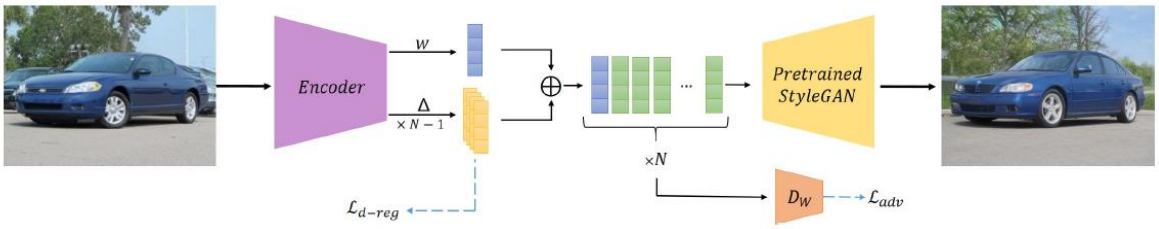
۴ – ۳- معماری کدگذاری برای ویرایش

معماری کدگذاری برای ویرایش[۲۵](به اختصار e4e) یک معماری برای نگاشت تصویر ورودی به فضای نهفته w با هدف اعمال ویرایش در تصویر می‌باشد که در پژوهش [۸] معرفی و ارائه شده است که اساس آن بر مبتنی بر معماری pSp [۵] بوده و ساختار آن در شکل (۹) قابل مشاهده است.

در بخش قبل ملاحظه شد که در پژوهش [۸] هدف نگاشت تصویر ورودی به w می باشد؛ روش پیشنهادی برای این مقصود کمینه کردن فاصله w و +w می باشد بدین نحو که اگر خروجی های e4e را w_0، w_1، و... فرض کنیم، مقصود کمتر

[25] Encoder For Editing

کردن فاصله‌ی هر یک از w_i ها با w می باشد طوری که در نهایت بردار های سبک حاصل بصورت رابطه (۳) پدید آید که در آن دلتا ها مقادیر اندک و قابل یادگیری می باشد(در شکل (۹) با رنگ زرد و آبی مشخص شده است).



شکل ٦: معماری کدگذاری برای ویرایش[۸]

$$E(x) = (w, w + \Delta_1, ..., w + \Delta_{N-1}) \qquad (۳)\ [۸]$$

فرآیند آموزش بدین صورت می‌باشد که ابتدا به تعداد تکرار محدودی با خطای برابری w_i و w آموزش e4e انجام شده(شرط همه دلتا ها برابر با صفر) و w تخمین زده می‌شود و سپس مقادیر دلتا ها یک به یک از شرط برابر صفر بودن خارج شده و هر کدام آنها طی چند تکرار یادگرفته می‌شود؛ برای آنکه دلتا ها مقادیر بزرگ به خود نگیرند، L_2 دلتاها به عنوان جمله منظم‌ساز به تابع هزینه افزوده می شود.

حال سوالی که مطرح می‌باشد این است که چگونه می‌توان اطمینان بیشتری حاصل کرد که w تخمینی در رنج w شبکه مولد StyleGAN بوده و از آن توزیع پیروی می‌کند؟ برای حل این موضوع از یک تمایزگر[۲۶] استفاده شده و خطای انتشاری از آن در قالب جمله منظم‌ساز باعث یادگیری این موضوع می‌شود که آیا w تخمینی در توزیع w شبکه مولد قرار می‌گیرد یا خیر (بلوک نارنجی در شکل (۹)). تابع هزینه برای آموزشی e4e نیز بصورت خود نظارت انجام شده و مجموع وزن‌دار سه شاخص می باشد که در رابطه (۴) آمده است؛ جمله مهم این تابع محاسبه فاصله دو تصویر ورودی و تصویر بازسازی شده توسط w تخمینی با شبکه مولد StyleGAN2 [۹] می باشد که تحت عنوان sim آمده است.

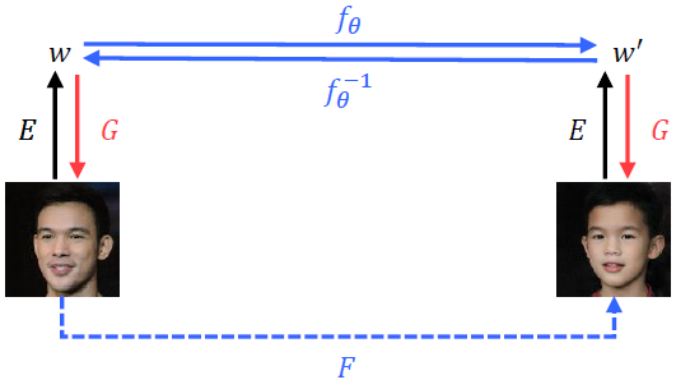
$$\mathcal{L}_{\text{dist}}(x) = \lambda_{l2}\mathcal{L}_2(x) + \lambda_{lpips}\mathcal{L}_{LPIPS}(x) + \lambda_{sim}\mathcal{L}_{\text{sim}}(x) \qquad (۴)\ [۸]$$

۴ – ۴- ارزیابی و نتایج تجربی آموزش شبکه کدگذاری برای ویرایش

آموزش شبکه e4e و ارزیابی آن بخاطر اینکه دارای ابعاد مختلف کمی و کیفی در بازسازی تصاویر و ویرایش‌پذیری آن می‌باشد، امری چالشی است لذا برای هر قسمت از آن شاخص و معیار هایی در نظر گرفته شده است که در ادامه توضیح داده خواهد شد:

[26] Discriminator

- بازسازی تصویر: برای ارزیابی اینکه e4e در چه حد توانایی بازسازی تصویر بدون اعمال ویرایش و بر اساس کد w تخمینی را دارد، از شاخص های L_2 و LPIPS استفاده می گردد.
- کیفیت ادراکی: برای اینکه ارزیابی کیفیت تصاویر بازسازی شده از دید ناظر و بر اساس درک بصری انجام شود، از فاصله های FID و Sliced Wasserstein Distance (به اختصار SWD) برای e4e استفاده شده است.
- ویرایش پذیری: برای اینکه بتوان امکان و سطح ویرایش تصویر در w تخمینی توسط e4e را ارزیابی نمود با استفاده از چند متد مبتنی بر سبک، w ویرایش شده بصورت معکوس به تصویر تبدیل شده و سپس توسط FID و SWD ویرایش انجام شده سنجش می شود.
- سازگاری ویرایش در فضای نهفته w: در پژوهش [۸] برای اینکه بتوان سازگاری و ثبات ویرایش در فضای پنهان را ارزیابی نمود، شاخص Latent Editing Consistency (به اختصار LEC) معرفی گردیده است؛ فرآیند این شاخص در شکل (۱۰) قابل مشاهده بوده و رابطه آن نیز در (۵) آمده است. هدف این شاخص بدین گونه می‌باشد که ابتدا تصویری ورودی را توسط e4e به فضای نهفته w نگاشت می‌کنیم و سپس با اعمال تغییر در آن، تصویر ویرایش شده را بدست آوردیم؛ حال مجددا تصویر ویرایش شده را توسط e4e به فضای نهفته w نگاشت کرده و معکوس ویرایش انجام شده را اعمال می‌کنیم؛ فاصله‌ی w حاصل با w قبل از ویرایش نشانگر شاخص LEC خواهد بود (فاصله بر اساس L_2).



شکل ۱۰: طرحواره شاخص سازگاری ویرایش در فضای نهفته[۸]

$$LEC(f_\theta) = \mathop{\mathbb{E}}_{x} \left\| E(x) - f_\theta^{-1}(E(G(f_\theta(E(x))))) \right\|_2 \qquad (۵)\ [۸]$$

حال برای ارزیابی تجربی و سنجش شاخص ها، چهار حالت با دو دامنه‌ی مختلف تصاویر خودرویی و چهره در نظر گرفته شده است که در جدول (۳) قابل مشاهده است. چهار حالت ترکیبی وجود یا عدم وجود جملات منظم ساز مورد اشاره در بخش (۳-۴) می باشد. (A) یعنی هیچ‌یک از جملات اعمال نشده است، (B) یعنی فقط جمله‌ی منظم‌ساز L_2 مربوط به کمینه کردن دلتاها اعمال شده است، (C) یعنی فقط جمله منظم‌ساز مربوط به تمایزگر اعمال شده است و (D) یعنی هر دو اعمال شده است.

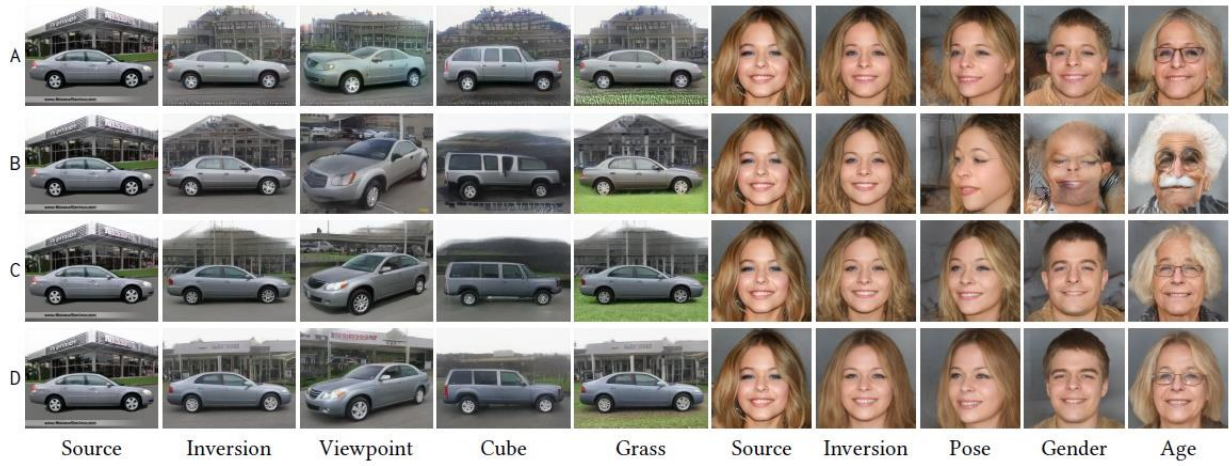
جدول ۳: نتایج شاخص های ارزیابی برای e4e [۸]

Domain	Conf.	Distortion		Perception		Editability	
		L_2	LPIPS	FID	SWD	FID	SWD
Faces	A	**0.03**	**0.17**	**25.17**	48.72	**62.46**	48.75
	B	0.04	0.18	28.09	39.98	149.85	69.86
	C	0.04	0.19	27.36	**33.96**	143.03	**41.94**
	D	0.05	0.23	30.96	40.54	81.08	43.63
Cars	A	**0.10**	0.32	**10.56**	**22.08**	**12.92**	**24.30**
	B	0.11	0.32	11.16	24.06	17.85	29.78
	C	0.11	0.33	12.47	38.08	16.22	44.73
	D	**0.10**	0.32	12.18	22.71	15.44	31.45

برای شاخص LEC نیز جدول (۴) در [۸] گزارش شده است؛ همانطور که قابل ملاحظه است در صورت اعمال هر دو جمله منظم ساز، سازگاری ویرایش در فضای نهفته بالاتر می‌رود. همچنین در شکل (۱۱) نمونه ویرایش های انجام شده با کدگذار e4e قابل مشاهده می‌باشد؛ بصورت واضح از تصاویر حاصل قابل درک است که در صورت اعمال توامان هر دو جمله منظم‌ساز مذکور در بخش (۴-۳)، تصاویر ویرایشی به جهت ادارکی با کیفیت بوده و ناظر شاید به سختی بتواند متوجه ویرایش انجام شده شود.

جدول٤: نتایج شاخص LEC برای e4e [۸]

	Faces				
Conf. f_θ	Young	Old	Smile	No Smile	Avg
A	63.03	59.42	48.08	48.15	54.67
D	**24.33**	**24.58**	**19.92**	**20.46**	**22.32**
	Cars				
Conf. f_θ	Pose I	Pose II	Cube	Color	Grass	Avg
A	186.95	181.50	133.93	83.93	89.78	135.22
D	**56.28**	**56.37**	**70.46**	**28.19**	**33.06**	**48.87**



شکل ۷: نمونه ویرایش های انجام شده در تصاویر توسط e4e در [۸]

فصل ۵ – قطعه‌بندی بی‌نظارت[۲۷] تصاویر مبتنی بر سبک

در دو فصل قبل در خصوص امکان ترجمه تصویر به تصویر و ویرایش معنی‌دار تصویر با استفاده از StyleGAN دو پژوهش [۵] و [۸] مورد بررسی قرار گرفته و نشان داده شد که با دسترسی و کنترل بر سبک های تولیدی و بردار های فضای نهفته‌ی یک تصویر می‌توان در آن تغییر مد نظری را ایجاد نمود؛ حال اگر هدف قطعه‌بندی اشیای موجود در تصویر تولیدی باشد، می‌توان با استفاده از بردار فضای نهفته‌ی آن و تغییر قسمت رنگ پیش‌زمینه[۲۸] و پس‌زمینه[۲۹] ماسک مربوط به تصویر قطعه‌بندی شده را خروجی گرفت. در مقاله [۱۰] هدف قطعه‌بندی تصویر تولیدی به صورت بی‌نظارت می‌باشد که یک معماری با عنوان Labels4Free (به اختصار l4f) پیشنهاد شده است که در این فصل مورد مطالعه قرار خواهد گرفت.

۵ – ۱– اساس قطعه بندی بی‌نظارت مبتنی بر سبک

قسمت مولد شبکه StyleGAN و خصوصا بردار های معادل در فضای نهفته w (بردار های سبک) دارای اطلاعات ارزشمندی در خصوص تصویر تولیدی و ویژگی های آن می‌باشد لذا این بردار ها برای قطعه‌بندی نیز مهم هستند چرا که قطعه‌بندی نیز بر اساس ویژگی های بصری و سبکی انتقالی به تصویر حاصل می‌شود (۱)؛ همچنین در فصل دوم ملاحظه شد که سطوح مختلف فرا نمونه برداری در شبکه مولد StyleGAN، تشکیل دهنده‌ی اجزای مختلف معنی‌دار تصویر تولیدی شامل پس‌زمینه، پیش‌زمینه و جزئیات آن دو می‌باشد لذا در صورت مستقل بودن ویژگی های پس‌زمینه و پیش‌زمینه در فضای نهفته w ، می‌توان ماسک قطعه بندی شده را بر اساس تغییر آنها ایجاد نمود(۲). توجه به این دو نکته اساس راهکار پیشنهادی در l4f برای تولید تصاویر مبتنی بر سبک به همراه ماسک قطعه بندی می‌باشد که در بخش (۵-۲) مورد بررسی قرار می‌گیرد.

۵ – ۲– ساختار و معماری شبکه قطعه بندی بی‌نظارت مبتنی بر سبک

در مقاله‌ی [۱۰] هدف اصلی تولید تصاویر مبتنی بر سبک می‌باشد به طوری ماسک معادل با قطعه‌بندی پیش‌زمینه و پس‌زمینه‌ی آن نیز حاصل شود؛ برای این منظور از دو شبکه مولد(حاصل از شبکه StyleGAN نسخه دوم [۹]) و پیش‌آموزش دیده، یک شبکه تمایزگر(آموزش دیده نشده) و یک شبکه با عنوان Alpha برای تعیین ماسک قطعه‌بندی استفاده شده است.

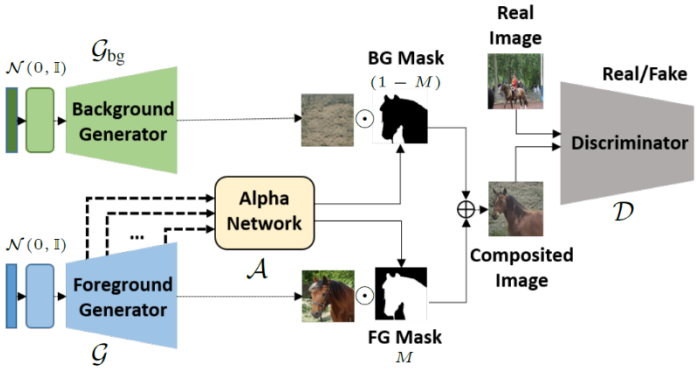
یکی از شبکه های مولد(G) بصورت عادی اقدام به تولید تصویر با شی مد نظر می‌کند، شبکه مولد دیگر (G_{gb}) اقدام به تولید تصویر پس زمینه خالی و بدون هیچ شی می‌کند، شبکه Alpha با استفاده خروجی های هر مرحله‌ی شبکه‌ی (G) اقدام به تولید یک ماسک دودوئی می‌کند. بر اساس ماسک حاصل، تصویر خروجی شبکه‌ی (G) و (G_{gb}) با هم ادغام شده و به

[27] Unsupervised

[28] Foreground

[29] Background

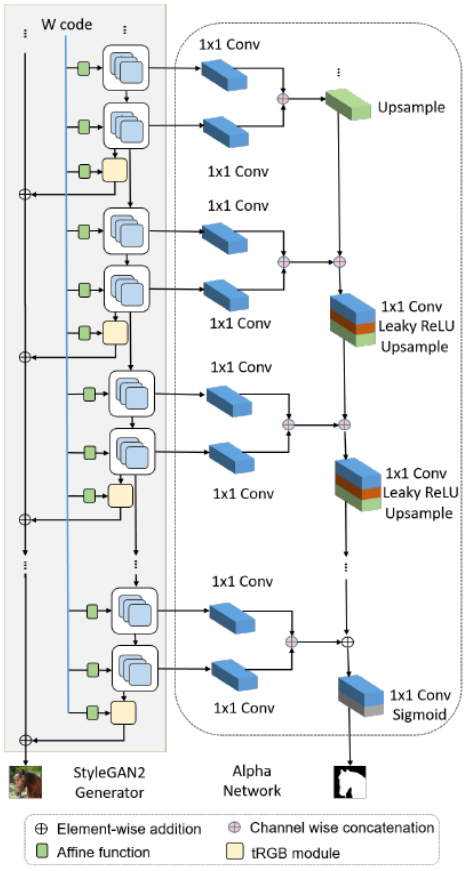
تمایزگر داده می‌شود؛ در فرآیند مذکور، شبکه‌ی Alpha و تمایزگر بصورت تقابلی با هم آموزش داده می‌شود. طرحواره این معماری در شکل (۱۲) قابل مشاهده است.



شکل ۱۲: ساختار Labels4Free برای تولید ماسک قطعه‌بندی شده [۱۰]

از جایی که تمایزگر از پیش آموزش دیده شده‌ی شبکه‌های StyleGAN بسیار قدرتمند بوده و فرصت کافی برای آموزش شبکه Alpha وجود ندارد، از آموزش دیده شده‌ی آن استفاده نمی‌شود. شکل (۱۳) معماری شبکه‌ی Alpha را نشان می‌دهد که با ورودی گرفتن از سطوح مختلف شبکه مولد (G)، اقدام به تولید ماسک قطعه‌بندی می‌کند؛ از جایی که صرفا تعیین پیکسل های تصویر بصورت دودوئی هدف می‌باشد، از لایه‌ی پیچشی با سایز ۱*۱ استفاده می‌شود تا تعداد صفحات ویژگی ضمن حفظ اطلاعات، کمتر شود. همچنین خروجی های سطح ابتدایی شبکه مولد به تنهایی امکان تصمیم‌گیری در خصوص تعیین ماسک دقیق را ندارد، لذا از آن فقط فرا نمونه برداری می‌شود تا برای مراحل بعد آماده شود؛ از مراحل بعد، خروجی ترکیب صفحات ویژگی قبل از فرا نمونه برداری از تابع فعالیت غیرخطی عبور داده می‌شود و در گام نهایی از تابع Sigmoid گذرانده می‌شود تا خروجی در بازه [۱-۰] حاصل شده و آماده‌ی تعیین ماسک شود. در حالت مذکور و در زمان پردازش، پیکسل های همسایه با هم تعامل ندارند که باعث می‌شود تعیین ماسک بر اساس ویژگی ها و سبک های انتقالی در پیکسل پردازشی حفظ شده و بر هم اثر نداشته باشند.

حال سوالی که مطرح است این می‌باشد که با چه تضمینی شبکه Alpha یک ماسک تمام فعال برای شبکه (G) تولید ننماید؛ برای پاسخ به این چالش از جمله های منظم‌ساز مبتنی بر تعداد پیکسل های فعال در تابع هزینه شبکه استفاده می‌شود تا تعداد آن کمینه باشد. طبق تحقیق و گزارش [۱۰] در شبکه (G_{gb}) نیاز داریم تصویر پس‌زمینه‌ای خالی از هر گونه شی تولید نماییم؛ برای این منظور لایه‌ی اول (G_{gb}) برابر با صفر در نظر گرفته می شود(این تصمیم بر اساس اعمال گرادیان بر خروجی لایه ها و محاسبه تاثیر آن در تصویر نهایی بدست آمده است).



شکل ۱۳: معماری شبکه **Alpha** در ساختار **Labels4Free** [۱۰]

۵ – ۳– آموزش و نتایج تجربی شبکه Labels4Free برای تولید تصویر قطعه‌بندی شده

آموزش و مقایسه l4f با چندین مجموعه داده‌ی مختلف(چهره، اسب و...)، با ساختار های مختلف(با نظارت یا بی‌نظارت) و با داده های ساختگی/واقعی در [۱۰] انجام و گزارش شده است که حاکی از عملکرد مطلوب آن می‌باشد. به جهت ساختار آموزش بی‌نظارت، رقیب مستقیم l4f [۱۰] شبکه PSeg [۱۲] می‌باشد؛ نتایج مقایسه این دو در جدول (۵) آمده است و قابل ملاحظه می‌باشد که در شاخص های مختلف، l4f توانسته است بین ۱۰٪ الی ۳۵٪ عملکرد بهتری داشته باشد.

در مقایسه‌ی جدول (۵) بین PSeg [۱۲] و l4f [۱۰] از شبکه‌ی BiSeNet [۱۴] که با مجموعه داده‌ی CelebA-Mask [۱۳] آموزش داده شده است، به عنوان مبنای صحت[۳۰] استفاده شده است.

جدول ۱: مقایسه معماری **PSeg** [۱۲] با **Labels4Free** [۱۰] در [۱۰]

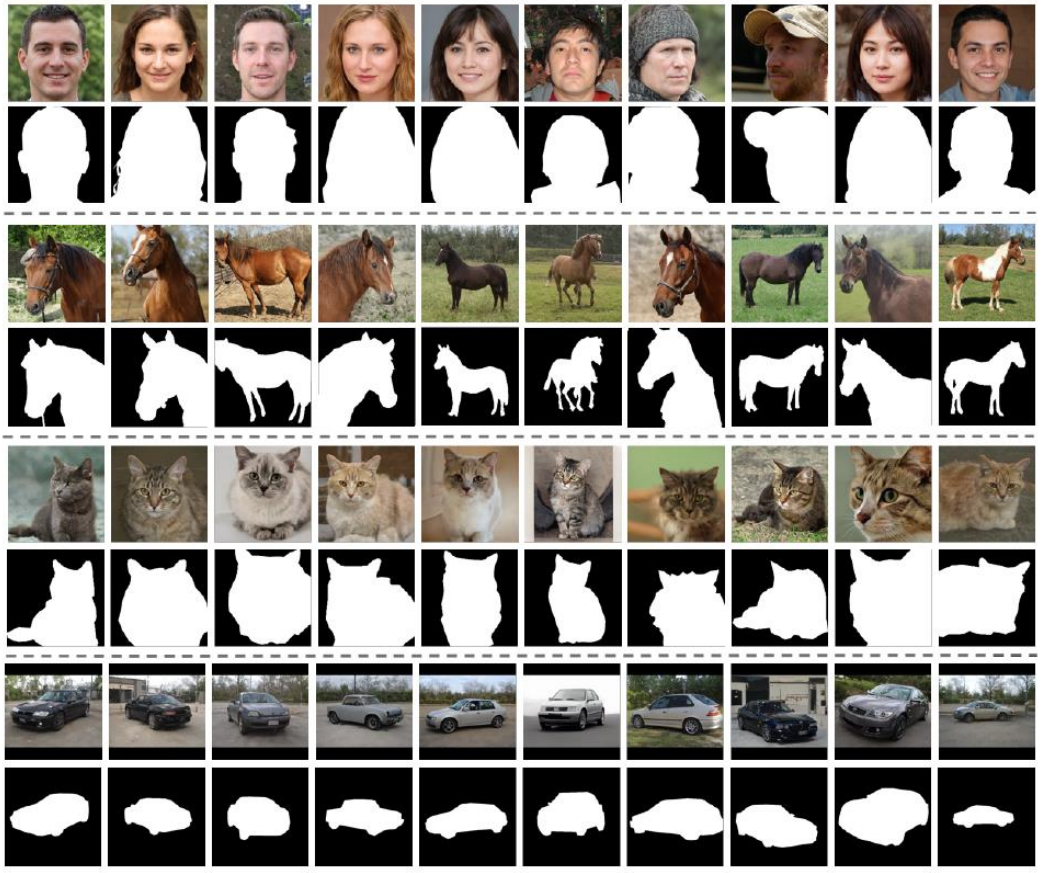
Method	Truncation Ψ = 0.7						Truncation Ψ = 1.0					
	IOU fg/bg	mIOU	F1	Prec	Rec	Acc	IOU fg/bg	mIOU	F1	Prec	Rec	Acc
PSeg	0.52/0.82	0.67	0.80	0.78	0.81	0.85	0.50/0.81	0.66	0.78	0.77	0.80	0.84
Ours	**0.87/0.94**	**0.90**	**0.95**	**0.95**	**0.94**	**0.95**	**0.75/0.89**	**0.82**	**0.90**	**0.92**	**0.89**	**0.92**

[30] Ground truth

پژوهش های انجامی در [۱۰] جهت ارزیابی l4f صرفا به مجموعه داده‌ی چهره CelebA-Mask [۱۳] محدود نبوده و با سایر مجموعه داده ها نظیر LSUN-Car، LSUN-Horse و LSUN-Cat [۱۵] مقایسه با PSeg [۱۲] انجام پذیرفته و در جدول (۶) قابل مشاهده بوده و حاکی از برتری دارد. همچنین در شکل (۱۴) نمونه هایی از قطعه‌بندی انجام شده با l4f قابل ملاحظه است.

جدول ۲ : مقایسه معماری PSeg [۱۲] با Labels4Free [۱۰] با سری مجموعه داده های [۱۵] در [۱۰]

Method	IOU fg/bg	mIOU	F1	Prec	Rec	Acc
$PSeg$(A)	0.65/0.68	0.66	0.80	0.81	0.80	0.79
Ours(A)	**0.84/0.77**	**0.81**	**0.89**	**0.89**	**0.90**	**0.90**
$PSeg$(B)	0.50/0.40	0.45	0.71	0.69	0.73	0.63
Ours(B)	**0.83/0.67**	**0.75**	**0.85**	**0.84**	**0.91**	**0.87**
$PSeg$(C)	0.81/0.73	0.77	0.83	0.83	0.84	0.85
Ours(C)	**0.93/0.84**	**0.89**	**0.94**	**0.93**	**0.95**	**0.95**



شکل ۸ : نمونه تصاویر قطعه بندی شده با l4f [۱۰]

فصل ۶ – بهبود شبکه مولد تقابلی مبتنی بر سبک و بررسی فضای سبک آن

در برخی از تصاویر تولیدی توسط شبکه StyleGAN عارضه های بصری به چشم می‌خورد که در صورت دقت ناظر انسانی قابل تشخیص می‌باشد؛ برای حل این مشکل و همچنین افزایش کیفیت تصاویر تولیدی و جزئیات آن، نسخه دوم شبکه StyleGAN با عنوان StyleGAN2 [۹] معرفی و پیشنهاد شده است. در بخش اول از این فصل StyleGAN2 بصورت کلی مورد معرفی قرار گرفته و در ادامه بررسی فضای سبکی و ویژگی های آن که حاصل پژوهش [۱۶] می‌باشد، گزارش خواهد شد.

۶ - ۱- نسخه دوم شبکه StyleGAN و تغییرات آن

در شبکه StyleGAN [۲] برخی ناهنجاری های غیرعادی بصری جزئی در تعدادی از تصویر تولیدی وجود دارد که نمونه هایی از آن در شکل (۱۵) قابل مشاهده است؛ StyleGAN2 [۹] با هدف رفع این گونه چالش ها و بهبود کیفیت کلی تصاویر پیشنهاد گردیده است.

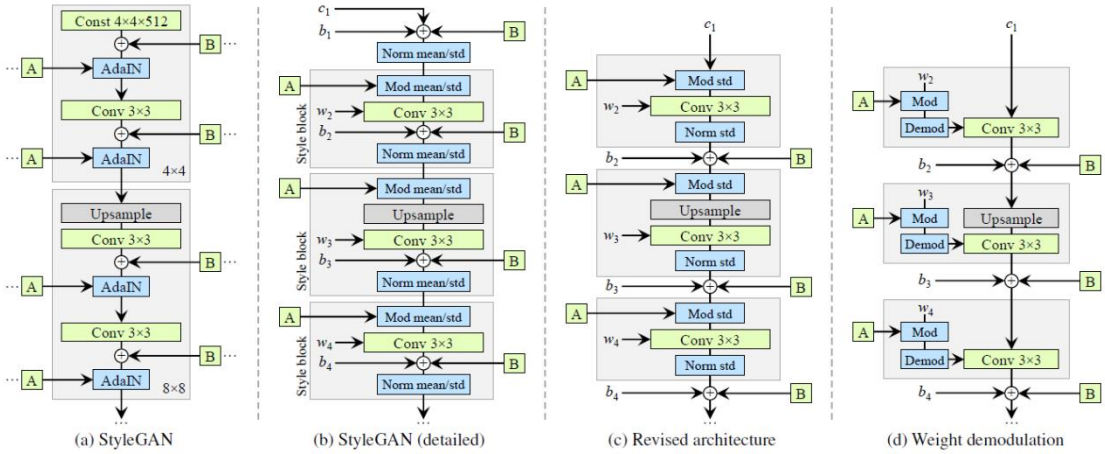
شکل ۹: عارضه های موجود در تصاویر تولیدی توسط شبکه StyleGAN [۹]

طی تحقیق انجام شده در [۹] مشخص شده است که عارضه های موجود در تصاویر تولیدی StyleGAN ناشی از اعمال محاسبات AdaIN در شبکه مولد می باشد؛ لذا به عنوان اولین تغییر اساسی، عملیات AdaIN با عملیات کشف وزن[31] برای انتقال سبک جایگذین شده است؛ در عملیات کشف وزن، ابتدا w با ضرایبی چون s مقیاس شده و سپس نرمال می‌شوند که روابط (۵) و (۶) نشان دهنده‌ی آن بوده و در شکل (۱۶) نیز ساختار آن قابل ملاحظه می باشد.

$$w'_{ijk} = s_i \cdot w_{ijk}, \qquad (۵)\ [۹]$$

$$w''_{ijk} = w'_{ijk} \Big/ \sqrt{\sum_{i,k} {w'_{ijk}}^2 + \epsilon}, \qquad (۶)\ [۹]$$

[31] Weight Demodulation



شکل ۱۰: ساختار شبکه StyleGAN2 که AdaIN به کشف وزن تبدیل شده است [۹]

به عنوان دومین تغییر در شبکه StyleGAN، گام اضافه کردن نویز و جانب[۳۲] به خارج از بلوک های انتقال سبک جا به جا شده است تا باعث شود اضافه کردن نویز در محاسبه آمارگان خروجی بلوک تاثیر نداشته و کیفیت تصویر خروجی افزایش یابد. نمونه تصاویر تولیدی حاصل از دو تغییر مذکور در شبکه [۹] در شکل (۱۷) قابل مشاهده است که عارضه های کوچک حذف شده‌اند.

شکل ۱۱: نمونه تصاویر تولیدی توسط شبکه StyleGAN2 که عارضه ها بصری در آنها حذف شده است.[۹]

یکی دیگر از عارضه هایی که در تصاویر تولیدی توسط StyleGAN وجود داشت، عارضه‌ی فازی بود که تغییر در جزئیات سطح پایین یا میانی تصویر که حاصل از تغییر در بردار w می‌باشد باعث تغییر در تصویر نهایی نمی‌شد؛ به عنوان مثال اگر جهت صورت تغییر پیدا می‌کرد، جهت اندام های جزئی چهره نظیر لب و دهن تغییر پیدا نمی‌کرد که نمونه این مورد در

[32] Bias

شکل (۱۸) قابل ملاحظه است. برای حل این مشکل از اتصال باقیماندگی[۳۳] در شبکه مولد و تمایزگر استفاده شده است تا بتوان تغییرات حاصل بین لایه ها انتقال داده شود. ایده‌ی این راه حل از پژوهش MSG-GAN [۱۷] بوده است که بتوان خطای حاصل را به تمامی سطوح شبکه‌ها رساند.

شکل ۱۲: نمونه عارضه‌ی فاز در تصاویر تولیدی توسط StyleGAN [۹]

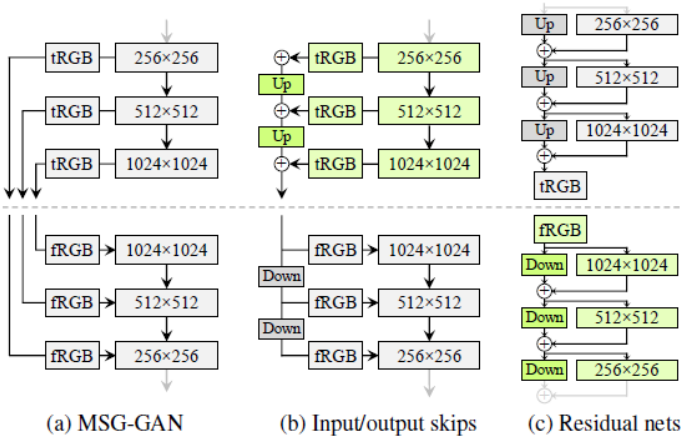
ساختار ایجاد اتصال باقیماندگی که در پژوهش [۹] پیشنهاد شده است، در شکل (۱۹) قابل ملاحظه است؛ همچنین علاوه بر ایجاد دو تغییر یاد شده در ساختار شبکه StyleGAN ، دو جمله‌ی منظم‌ساز برای اضافه کردن به تابع هزینه پیشنهاد شده است که اهداف آن افزایش کلی کیفیت(وضوح مکانی) و هموار ساختن جزئیات تصاویر تولیدی توسط w می‌باشد که در این بخش به آنها پرداخته نمی‌شود.

علاوه بر رفع عارضه های بصری در تصاویر تولیدی که بصورت شهودی نمایان است، این امر در گزارش شاخص های کمی نظیر FID نیز قابل ملاحظه است که در جدول (۷) آمده است.

جدول ۳ : مقایسه شبکه StyleGAN و StyleGAN2 [۹]

Configuration	FFHQ, 1024×1024				LSUN Car, 512×384			
	FID ↓	Path length ↓	Precision ↑	Recall ↑	FID ↓	Path length ↓	Precision ↑	Recall ↑
A Baseline StyleGAN [24]	4.40	212.1	**0.721**	0.399	3.27	1484.5	**0.701**	0.435
B + Weight demodulation	4.39	175.4	0.702	0.425	3.04	862.4	0.685	0.488
C + Lazy regularization	4.38	158.0	0.719	0.427	2.83	981.6	0.688	0.493
D + Path length regularization	4.34	**122.5**	0.715	0.418	3.43	651.2	0.697	0.452
E + No growing, new G & D arch.	3.31	124.5	0.705	0.449	3.19	471.2	0.690	0.454
F + Large networks (StyleGAN2)	**2.84**	145.0	0.689	**0.492**	**2.32**	**415.5**	0.678	**0.514**
Config A with large networks	3.98	199.2	0.716	0.422	–	–	–	–

[33] Residual



شکل ۱۹: اضافه کردن ساختار باقیماندگی به شبکه های مولد و تمایزگر StyleGAN [۹] – نیمه‌ی بالایی مربوط به شبکه مولد و نیمه‌ی پایینی مربوط به شبکه تمایزگر می‌باشد.

۶ – ۲- فضای نهفته‌ی سطح بالاتر و تغییرپذیری بهتر

در چهار فصل قبل ملاحظه کردیم که برای هر مسئله‌ای، داشتن فضای نهفته‌ای که برای ما تفسیرپذیرتر بوده و تغییرات کنترل شده در آن باعث ایجاد تغییرات مستقیم و قابل پیش‌بینی در تصویر نهایی باشد، ارزشمند است. از این رو، مهم است که متغیر های فضای نهفته به یکدیگر وابسته نبوده و تغییر در هر یک بتواند تغییر بخصوصی در تصویر تولیدی ایجاد نماید(ویژگی گسیختگی[۳۴] متغیر ها). همچنین معکوس رابطه فوق نیز از اهمیت بسزایی برخوردار است بطوری که برای هر تغییر سطح بالا در تصویر نهایی، صرفا یک (یا اندک) متغیر از فضای نهفته بتواند آن تغییر را ایجاد نماید(ویژگی کامل بودن[۳۵] متغیر ها).

در StyleGAN2 [۹] و در بخش ۶-۱ ملاحظه کردیم که اعمال فضای نهفته w به عنوان سبک انتقالی به تصویر ورودی با استفاده از متد کشف وزن باعث بهبود کیفیت تصاویر ضمن حذف برخی از عارضه های بصری می‌شود؛ این خود فرآیندی جهت تبدیل متغیر های فضای نهفته w به فضای سبکی مبتنی بر کانال(به اختصار StyleSpace) می‌باشد؛ حال سوالی که مطرح می‌شود این است که StyleSpace تا چه حد پاسخگوی ویژگی های گسیختگی و کامل بودن متغیرها می‌باشد. پاسخ به این سوال و بررسی فضای StyleSpace در بخش ۶-۳ مورد بحث و مطالعه قرار خواهد گرفت.

۶ – ۳- انتخاب فضای نهفته مناسب

در StyleGAN2 ما با سه فضای نهفته‌ی z، w و StyleSpace(S) در روند تولید تصاویر در شبکه مولد مواجه هستیم که در پژوهش [۱۶] برای بررسی ویژگی های گسیختگی، کامل بودن و دارای اطلاعات مفید بین این سه فضا، آزمایشی صورت گرفته و معیار DCI [۳۶] [۱۱] برای آن گزارش شده است که در جدول (۸) قابل مشاهده است. در این آزمایش ابتدا ۴۰

[34] Disentanglement

[35] Completeness

[36] Disentanglement /Completeness/Informativeness

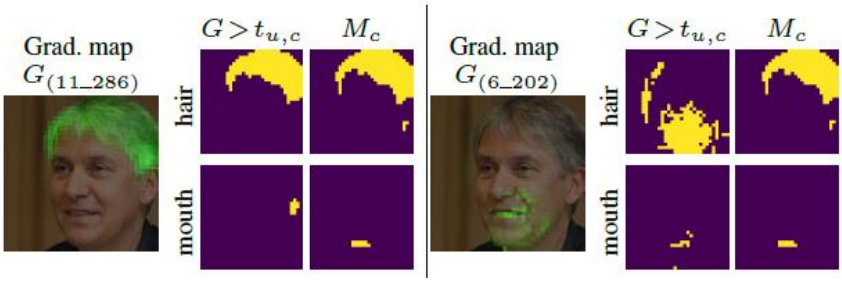
دسته‌بند دودوئی با مجموعه داده‌ی [۱۳] وجود ویژگی های چهره نظیر داشتن موی سفید، وجود لبخند بر لب و ... را آموزش می‌بینند تا سنجش DCI توسط نتایج آن انجام شود؛ سپس ۵۰۰,۰۰۰ تصویر ساختگی توسط StyleGAN2 تولید شده و توسط دسته‌بند های مذکور ارزیابی شده و گزارش نهایی حاصل می‌شود. همانگونه که در جدول (۸) مشهود است متغیر های فضای نهفته‌ی S نسبت به فضا های z و w دارای از هم گسیختگی بیشتری بوده و برای کنترل بر ویژگی های سطح بالای تصویر تولیدی انتخاب بهتری می‌باشد.

جدول ٤: مقایسه فضا های نهفته z، w و S [١٦]

	Comparison w/ $\mathcal{Z}$ and $\mathcal{W}$				Comparison with $\mathcal{W}+$			
	Disent.	Compl.	Inform.		Disent.	Compl.	Inform.
$\mathcal{Z}$	0.31	0.21	0.72				
$\mathcal{W}$	0.54	0.57	0.97	$\mathcal{W}+$	0.54	0.64	0.94
$\mathcal{S}$	**0.75**	**0.87**	**0.99**	$\mathcal{S}$	**0.63**	**0.81**	**0.98**

۶ - ۴- کانال های فعال فضای StyleSpace بر ویژگی های محلی معنادار

در آزمایش دیگری که در پژوهش [۱۶] و برای تحلیل فضای StyleSpace صورت گرفته است، هدف شناسایی کانال هایی از StyleSpace می‌باشد که با تغییر آن، ناحیه‌ای محلی و معنادار از تصویر متحول تغییرات می‌شود(و نه کل تصویر). برای این هدف ابتدا یک شبکه BiSeNet [۱۴] آموزش داده شده و برای بدست آوردن نواحی معنایی تصویر ورودی از آن استفاده می‌شود؛ سپس گرادیان تصویر به ازای هر کدام از کانال های StyleSpace محاسبه و با هر کدام از نقشه های معنایی در مرحله قبل مقایسه می‌شود تا نواحی همپوشانی شده با یکدیگر استخراج شود. این فرآیند به ازای تعدادی تصاویر تولیدی انجام می‌شود و در صورتی که همپوشانی ناحیه‌ی معنایی بخصوصی با کانال مشخصی از یک حد آستانه [۳۷] بیشتر شود، آن کانال به عنوان کانال فعال برای آن ناحیه‌ی محلی معنادار انتخاب می‌شود. نمونه‌ی این عملیات در شکل (۲۰) قابل ملاحظه است که کانال فعال مربوط به ناحیه‌ی موی سر و دهان شناسایی شده است.



شکل ۲۰: کانال های فعال فضای StyleSpace بر ویژگی های محلی معنادار دهان و موی سر [١٦]

[37] Threshold

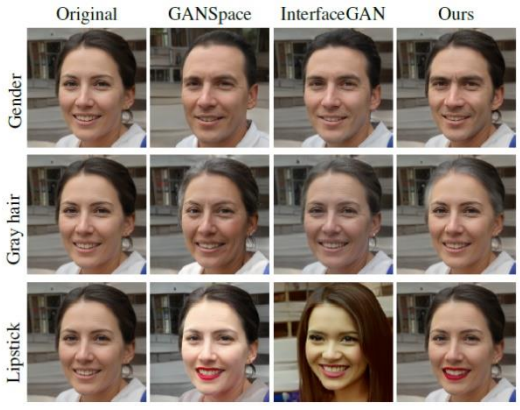
۶ - ۵- شناسایی کانال های فعال مربوط به یک ویژگی خاص

در صورتی هدف ما شناسایی کانال بخصوصی از فضای StyleSpace باشد بطوری که فعالیت آن کانال در ناحیه‌ی ویژگی مد نظر ما قرار داشته باشد، انجام آزمایشی شبیه آزمایش بخش ۶-۴ بسیار پر هزینه و شاید ناممکن است؛ به عنوان یک مثال برای آزمایش مذکور ممکن است هدف ما پیدا کردن کانال هایی از فضای StyleSpace باشد که خاکستری شدن موی سر را دنبال می‌کند. راه حل پیشنهادی برای این چالش در پژوهش [۱۶] استفاده از یک تکنیک ساده آماری و مبتنی بر توزیع است بدین نحوه که ابتدا چند نمونه‌ی محدود(مثلا ۱۰ الی ۳۰) که دارای ویژگی مد نظر می‌باشد را در نظر می‌گیریم(مجموعه‌ی مثبت ورودی)؛ سپس برای هر یک از نمونه‌های ورودی، فاصله‌ی نرمال شده‌ی بردار سبک آن با میانگین توزیع تصاویر تولیدی محاسبه می‌شود(δ^e) و سپس اختلاف مقدار حاصل با میانگین و انحراف معیار مجموعه‌ی ورودی مثبت بدست می آید(σ^e و μ^e)؛ حال به ازای هر کدام از کانال های سبک که نسبت اندازه μ^e به σ^e برای آن بیشینه باشد، نشان می‌دهد که کنترل ویژگی مد نظر توسط این کانال محتمل تر است.

۶ - ۶- ویرایش پذیری و دستکاری بهتر

در چهار بخش قبل و بر اساس پژوهش [۱۶] ملاحظه شد که با بکارگیری برخی تکنیک ها و با بهره از شبکه های جانبی نظیر BiSeNet [۱۴] می‌توان استخراج نمود که چه کانال هایی از StyleSpace در StyleGAN2 کنترل کننده‌ی چه ویژگی‌ای در تصویر نهایی می‌باشد(بر اساس گرادیان) یا برای ویژگی خاص، چه کانال هایی آن را کنترل می‌کند(بر اساس آمارگان توزیعی)؛ لذا امر ویرایش معنادار یا کنترل المان در تصاویر تولیدی می‌تواند با کیفیت و دقیق تر از قبل حاصل شده و دستکاری مد نظر در تصاویر تولیدی را بهتر از قبل ایجاد نمود؛ همچنین برای عکس های واقعی نیز با استفاده از تخمین فضای نهفته w و یا StyleSpace مربوطه، ویرایش مد نظر در تصویر را ایجاد نموده و تغییر یافته‌ی آن را مجددا بازسازی نمود.

با توجه به اینکه ویژگی های گسیختگی و کامل بودن در متغیر های StyleSpace از سطح بالایی برخوردار هستند، دستکاری های اعمالی در آن نیز نیز با کیفیت‌تر از قبل حاصل خواهد شد. برخی از نمونه های اعمال ویرایش و دستکاری با استفاده از فضای StyleSpace در شکل (۲۱) آورده شده است.



شکل ۱۳: نمونه ویرایش های اعمالی در StyleSpace [۱۶]

فصل ۷ – جمع بندی و مراجع

۷ - ۱- بحث

درک و شناخت پایه‌ی هر نوع مجموعه داده‌ای منجمله توزیع، متغیر های سازنده و... میتواند زمینه ساز حل مسائل گوناگونی در آن حیطه شود که در این گزارش داده های تصویری مورد بحث قرار گرفته و ملاحظه شد که تصاویر را می توان در ترکیب چندین سبک مختلف در سطوح متفاوت دانست.

در بحث تولید تصاویر ساختگی، شبکه های مولد تقابلی [۱] گام ارزشمندی برداشته است اما محدودیت آن در کنترل بر المان و ساختار تصاویر تولیدی ضعف مهمی تلقی می‌شد؛ با معرفی شبکه های مولد تقابلی مبتنی بر سبک [۲] و طراحی مجدد معماری شبکه‌ی مولد آن، مشکل مذکور حل شده و امکان کنترل بر سبک تصاویر تولیدی در سطوح مختلف معنایی مهیا شد. اینکه بتوان بر اساس بردار های فضای نهفته و سبک انتقالی، ویژگی های تصویر نهایی را کنترل و تغییر داد، دست‌آورد مهمی می‌باشد و از این نکته می‌توان برای پاسخگویی به مسائل مختلف حوزه تصویر استفاده نمود.

در پژوهش [۵] ملاحظه شد که با استفاده از فضای نهفته w در StyleGAN و حتی نگاشت تصاویر واقعی به آن با استفاده از هرم ویژگی و شبکه های عصبی پیچشی، می‌توان به مسائل مختلف حوزه ترجمه تصویر به تصویر نظیر افزایش وضوح مکانی، بازسازی، رنگ آمیزی تصویر، ترسیم چهره از رو به رو و ... پاسخ داد؛ لذا استفاده از فضای نهفته‌ی w برای ایجاد تصاویر جدید از اهمیت بسزایی برخوردار بوده و مطالعه‌ی آن زمینه‌ی تحقیقاتی فعالی محسوب می‌شود.

در مقاله [۸] مشاهده کردیم که تخمین فضای w به جای w^+ به ازای تصاویر واقعی می‌تواند امکان ویرایش هایی با دقت بیشتر را فراهم کند؛ همچنین در نتایج تحقیق [۱۰] نیز ملاحظه کردیم که با استفاده از دو شبکه مولد، می‌توان تصاویر پس‌زمینه و پیش‌زمینه را مجزا از هم تولید کرده و سپس ادغام نمود تا بدین صورت اشیای موجود در تصویر را قطعه‌بندی کرد که این متد مزیت های گوناگونی نظیر آموزش خودنظارتی را فراهم می‌کند. همچنین در نسخه دوم شبکه های مولد مبتنی بر سبک[۹] نیز مشاهده کردیم که برخی عارضه های بصری در تصاویر تولیدی رفع شده و کیفیت کلی افزایش یافته است و اعمال کشف وزن به جای عادی‌سازی تطبیقی نمونه، زمینه‌ساز یک فضای نهفته‌ی جدید با عنوان StyleSpace شده است که در [۱۶] مورد تحلیل گرفته و نشان داده شده است که می توان با تکنیک ها و آزمایش های ساده، کانال های سبکی را استخراج نمود که مستقیما ویرایش و تغییر مد نظر ما را در تصویر تولیدی ایجاد می‌کند و این حتی به تصاویر واقعی قابل بسط بوده و می‌توان با تخمین StyleSpace مربوطه، دستکاری مفروض را اعمال کرد.

۷ - ۲- فهرست مراجع

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

[2] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[3] Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision (pp. 1501-1510).

[4] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[5] Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2287-2296).

[6] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[7] 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.

[8] Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG), 40(4), 1-14.

[9] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119).

[10] Abdal, R., Zhu, P., Mitra, N. J., & Wonka, P. (2021). Labels4free: Unsupervised segmentation using stylegan. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13970-13979).

[11] Eastwood, C., & Williams, C. K. (2018, February). A framework for the quantitative evaluation of disentangled representations. In International Conference on Learning Representations.

[12] Bielski, A., & Favaro, P. (2019). Emergence of object segmentation in perturbed generative models. Advances in Neural Information Processing Systems, 32.

[13] Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5549-5558).

[14] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 325-341).

[15] Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365.

[16] Wu, Z., Lischinski, D., & Shechtman, E. (2021). Stylespace analysis: Disentangled controls for stylegan image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12863-12872).

[17] Zhang, F., & Wang, C. (2020). MSGAN: generative adversarial networks for image seasonal style transfer. IEEE Access, 8, 104830-104840.

[18] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586-595).

[19] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4690-4699).