


# Masking Cascaded Self-Attentions for Few-Shot Font-Generation Transformer

## — *Supplementary Material*

Jing Ma, Xiang Xiang\*<sup>(✉)</sup> , and Yan He

National Key Lab of Multi-Spectral Information Intelligent Processing Technology,  
School of Artificial Intelligence and Automation,  
Huazhong University of Science and Technology, Wuhan, China  
`xex@hust.edu.cn`

## 1 Appendix

In the appendix, we present the generated font images of unseen styles by FGTr for an ancient Chinese poem under a 10-shot setting; please refer to Fig. 2. These stylistic differences are manifested in the thickness, transition, and connection of strokes, and the spatial arrangement of components and radicals.

Additionally, we employ FGTr to produce a longer piece of traditional Chinese educational material, encompassing 1416 characters and 118 styles—including seen and unseen—randomly selected from our collected dataset; please refer to Fig. 3 and 4.

We provide network architectures of Style/Content Encoder and Decoder of our FGTr in Table. 2 and Table. 3, as a supplement to the implementation details in manuscript. Please note that the code utilized in this study is subject to copyright restrictions and, as such, cannot be made publicly available at the moment. We apologize for any inconvenience that may cause and appreciate your understanding.

### 1.1 Discussion

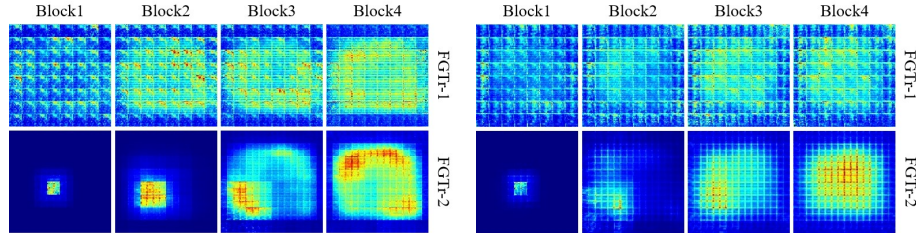
**How LSAM works?** We visualize the Effective Receptive Fields (ERF) of FGTr-1 (without LSAM) and FGTr-2 (with LSAM). The receptive field of four blocks is calculated using the absolute value of the gradient of the feature map’s center location to the input. Results are averaged across all channels in each map for 128 randomly selected font images. According to the visualization results in Fig. 1, LSAM enhances the attention between adjacent patches.

**Effectiveness of AGM?** We store content features  $I_c$  of real font images in the training dataset for content similarity scoring. When only one style-reference image is available, AGM still uses the confidence  $S_{conf}$  and content similarity  $S_{sim}^c$  to score multiple generated glyphs (by inputting the unique style-reference image  $I_s$  and multiple content images  $\{I_c\}_n$ ) and select the best one.

---

\* Also with Peng Cheng Laboratory, Shenzhen, China.

**Runtime of FGTr and baselines?** We apply our proposed FGTr and the baselines to generate 10,000 font images online and record the average time taken for each image. The results are shown in Table 1. Although FGTr contains many self-attention layers, the speed of self-attention operators is not slow. Therefore, FGTr demonstrates a good balance between runtime and generation quality.



**Fig. 1.** Effective Receptive Fields of FGTr-1 (w/o LSAM) and FGTr-2 (w/ LSAM). (Left) Style Encoder. (Right) Content Encoder.

**Table 1.** Average runtime of FGTr and baselines over generating 10,000 font images.

Runtime	DG-Font	CF-Font	FGTr
seconds (per image)	0.032	0.014	0.023

故人西辞黄鹤楼，	故人西辞黄鹤楼，
烟花三月下扬州。	烟花三月下扬州。
孤帆远影碧空尽，	孤帆远影碧空尽，
唯见长江天际流。	唯见长江天际流。
故人西辞黄鹤楼，	故人西辞黄鹤楼，
烟花三月下扬州。	烟花三月下扬州。
孤帆远影碧空尽，	孤帆远影碧空尽，
唯见长江天际流。	唯见长江天际流。
故人西辞黄鹤楼，	故人西辞黄鹤楼，
烟花三月下扬州。	烟花三月下扬州。
孤帆远影碧空尽，	孤帆远影碧空尽，
唯见长江天际流。	唯见长江天际流。
故人西辞黄鹤楼，	故人西辞黄鹤楼，
烟花三月下扬州。	烟花三月下扬州。
孤帆远影碧空尽，	孤帆远影碧空尽，
唯见长江天际流。	唯见长江天际流。
故人西辞黄鹤楼，	故人西辞黄鹤楼，
烟花三月下扬州。	烟花三月下扬州。
孤帆远影碧空尽，	孤帆远影碧空尽，
唯见长江天际流。	唯见长江天际流。

Fig. 2. 10-shot font generation results for an ancient Chinese poem with 10 different unseen styles using FGTr.

专插为仪知万顺中亥北名素走饲识嗅协伦忠终文原书义治求奥备恶事  
 以俱何礼先而妇乎至东之容下所所洞之则服说知四仁平讲之乐善其  
 贵名老习宜千夫应子在山不能人目鼻宜人臣五讲能至说至当书礼剔记  
 道子学友长千柔方支华岳常物畜色臭声族敬麻字说终德并径命言延要  
 之五不帅于而子四二此中五五五五五五五五五五五五五五五五五五五五  
 教教幼亲梯百父此十我此此此此此此此此此此此此此此此此此此此此此  
 迂方宜时梨百义东癸极衡信兽豕百朽入曾朋功遵繁初止子秋造记作子  
 乃义所少让而臣西至暖恒智鸟犬黑腥去玄与小共其有篇曾春训礼秋读  
 性有非方能十君曰甲温嵩礼有鸡及及曰至友大人惧必七乃礼有注春方  
 教山学子岁十者北者下华又鱼羊黄香上孙序衰学学者者学易模戴亡明  
 不燕不人四而纲南于道岱仁虫牛赤焦平子幼齐书广学子大书典小既既  
 苟冥子为融一三曰十赤曰日有马青臙曰自长斩惟若为孟作诗有大诗经  
 远抒情义拙文星穷数叔纪良咄食具含音孙恭背具乱读言易读详体咏梁  
 相机之知当某河不乎中之水所情所八而则违不可句善不可易治调谷  
 习断师不所识日运本当水国遍人七口乃子弟勿今不明记庸始三存当有  
 近学严学亲教者时行道读民物谷欲味竹子友叙艺继诂子偏经易官诗氏  
 相不不于某光四五赤四四植六恶五与而则师六草训弟不六周六四左  
 性子教人孝知三此此此此此此此此此此此此此此此此此此此此此此此此  
 善处过器席闻人冬土疆济商木褻惧咸金身从同数篆究篇笔熟藏礼颂羊  
 本邻之成温见地秋金所雅工草黍哀辛石而妇所书小讲十思书归周雅公  
 性择父不能次天日木日日有麦日及木父夫人御大颂二子四有作日有  
 初母教姆龄悌者夏火道河农生菽怒甘草祖恩义射文蒙者庸通山公风者  
 之孟不不九孝才春水黄江士所梁喜苦土曾子十乐古训语中经连周国传  
 人昔养玉香首三日曰曰曰曰地福日酸匏高父此礼有凡论作孝有我日三

Fig. 3. 10-shot font generation results (Part I) for a traditional Chinese educational material using FGTr, consisting of 1416 characters and 118 seen and unseen styles.

始世社久出篡晋齐基由世陵焚奇毁衰鉴斯勤苦卓思志之警是物后力  
 终盛夏长雄莽两高国有宋金器绩清兴通于且勤苦早立效自若如于勉  
 知称迂最七王造与创皆绝都神治满知参夕学自犹苦宜当亦不裕宜  
 系逊载载强平国周乱代金武出富帝乱经斯仕教劳生学于者学前哉  
 世指百霸孝三文隋五灭洪闾安允治证于既不虽小幼男为不于之  
 考相四八五至号字除称元号李氏传载兼朝彼彼身尔尔奇敏仕蜜母益  
 史帝下约国建鼎西师周帝明林嘉弱兹四惟论股角迟异奇敏仕蜜母益  
 诸二天诛战业汉东义汉称大如乾统在志而鲁刺挂悔称称聪已酿父无  
 读号家始终汉争分起及皆国寇历宣全国心读锥如犹众人且身蜂显戏  
 通虞子王秋兴吴魏祖晋金兴肆雅后史三涌令梁薪老成悟于幼丝声功  
 子有传武春祖蜀元高唐与祖淘康光今汉而中悬负既颖女虽吐名有  
 经唐夏周始高魏北唐梁辽太叔由同古后口赵头如彼彼彼晏蚕扬勤  
 庄世王亡说争献陵绪改混废禎定鄙国二目学勉辍籍士棋吟字人民经  
 老上三紂游汉于金统乃北祚崇大都民书亲勤知不书多赋咏正为泽一  
 及居称至尚楚终都失国南国至克扰建汉若尚且学读魁能能作曷下唯  
 子皇王戟戈世年朝传之传年世方法法一今贤书贫愤廷岁温童学君子  
 中三武百干二百年南再灭八十六四英宪记古圣无虽发大七道神不致教  
 文号周六逞传四为不梁十九十靖始立史通古彼家始对泌谢举苟上我  
 扬帝汤商坚并汉承字载禅代京命起制次录棠简雪七十二诗琴岁晨行赢  
 荀菁有号纳兼东陈土百周前燕景乱帝有实项竹映十咏辨七司而而满  
 有至商国王始为梁一三受超迁膺变废读考师荆如二八能能方鸡杜金  
 者农禹夏东氏兴继隋传兴广祖祖间兴繁者尼编萤泉灏岁姬晏夜学子  
 子羲有伐辙秦武齐至十宋图成世咸命虽史仲蒲囊老梁八文刘守而而遗  
 五目夏汤固赢光宋迨二炎輿迨清道革史读昔披如苏若莹蔡唐犬幼而遗

Fig. 4. 10-shot font generation results (Part II) for a traditional Chinese educational material using FGTr, consisting of 1416 characters and 118 seen and unseen styles.

Modules	Operation	Components
Input Image	Input	Resize, Normalize
Patch Embedding	Embedding	Reshape, Conv2d, Position
$2 \times$ Encoder Block	Attention MLP	LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Downgrade Layer	Downgrade	Reshape, LayerNorm, Linear
$2 \times$ Encoder Block	Attention MLP	LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Downgrade Layer	Downgrade	Reshape, LayerNorm, Linear
$6 \times$ Encoder Block	Attention MLP	LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Downgrade Layer	Downgrade	Reshape, LayerNorm, Linear
$2 \times$ Encoder Block	Attention MLP	LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Classifier	Classify	LayerNorm, Linear

**Table 2.** Network architecture of Style/Content Encoder. An encoder contains four groups of Blocks, with quantities of [2, 2, 6, 2], and the dimensions of the activation values are [192, 384, 768, 1536], respectively. A Downgrade Layer is inserted between two groups of Blocks for dimensionality increase.

Modules	Operation	Components
Sequences Concat	Concat	Concat
$3 \times$ Merging Block	Attention MLP Skip-Connection	LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear LayerNorm, Linear
$2 \times$ Decoder Block	Attention Attention MLP	LayerNorm, Cross-Attention(LSAM) LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Upgrade Layer	Upgrade	Reshape, LayerNorm, Linear
$6 \times$ Decoder Block	Attention Attention MLP	LayerNorm, Cross-Attention(LSAM) LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Upgrade Layer	Upgrade	Reshape, LayerNorm, Linear
$2 \times$ Decoder Block	Attention Attention MLP	LayerNorm, Cross-Attention(LSAM) LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Upgrade Layer	Upgrade	Reshape, LayerNorm, Linear
$2 \times$ Decoder Block	Attention Attention MLP	LayerNorm, Cross-Attention(LSAM) LayerNorm, Self-Attention(LSAM) LayerNorm, Linear, GELU, Linear
Output Image	Output	LayerNorm, Linear, Tanh

**Table 3.** Network architecture of Decoder, which uses the concatenated style and content sequences as input. Three Merging Blocks’ output activation dimensions are [3072, 1536, 1536]. A decoder contains four groups of Blocks, with quantities of [2, 6, 2, 2], and the dimensions of the activation values are [1536, 768, 384, 192], respectively. An Upgrade Layer is inserted between two groups of Blocks for dimensionality reduction.