

U-GAT-IT: 带有自适应层实例归一化的未监督的注意力生成网络, 用于图像到图像的翻译

Junho Kim^{1,2*}, Minjae Kim², Hyeonwoo Kang², Kwanghee Lee^{3 †}

¹Clova AI Research, NAVER Corp, ²NCSoft, ³Boeing Korea Engineering and Technology Center

jhkim.ai@navercorp.com, fminjaekim, hwkang0131g@ncsoft.com, kwanghee.lee2@boeing.com

摘要

我们提出了一种无监督的图像到图像翻译的新方法, 该方法以端到端的方式结合了新的注意力模块和新的可学习的归一化函数。注意力模块根据辅助分类器获得的注意力图, 将模型聚焦于区分源域和目标域中的更重要区域。与以前的基于注意力的方法无法处理域之间的几何变化不同, 我们的模型可以转换需要整体变化的图像和需要大范围形状变化的图像。此外, 我们新的 AdaLIN (自适应层实例归一化) 函数可帮助我们的注意力指导模型根据数据集灵活地控制学到的参数来控制形状和纹理的变化量。实验结果表明, 与现有的具有固定网络架构和超参数的最新模型相比, 我们的方法具有优越性。我们的代码和数据集可在以下位置获得: <https://github.com/taki0112/UGATIT> 或 <https://github.com/znxlw/UGATIT-pytorch>。

1 介绍

图像到图像的翻译旨在学习一种在两个不同域中映射图像的功能。由于该主题的广泛应用, 包括图像修复 (Pathak 等人 (2014); lizuka 等人 (2017)), 超分辨率 (Dong 等人 (2016); Kim 等人 (2016)), 图像着色 (Zhang 等人 (2016; 2017)) 和样式转换 (Gatys 等人 (2016); Huang & Belongie (2017)) 等, 并因此在机器学习和计算机视觉领域受到了研究人员的广泛关注。当给出配对样本时, 可以使用条件生成模型 (Isola 等人 (2017); Li 等人 (2017a); Wang 等人 (2018)) 或简单回归模型 (Larsson 等人 (2016); Long 等人 (2015); Zhang 等人 (2016)) 以监督方式训练映射模型。在没有配对数据的无监督环境下, 众多的工作 (Anoosheh 等人 (2018); Choi 等人 (2018); Huang 等人 (2018); Kim 等人 (2017); Liu 等人 (2017); Royer 等人 (2017); Taigman 等人 (2017); Yi 等人 (2017); Zhu 等人 (2017)) 已成功使用共享潜在空间翻译了图像 (Liu 等人 (2017)) 和周期一致性假设 (Kim 等人 (2017); Zhu 等人 (2017))。这些工作得到了进一步发展, 以处理多种形式的任务 (Huang 等人 (2018))。

尽管取得了这些进步, 但先前的方法仍显示出性能差异, 具体取决于域之间形状和纹理的变化量。例如, 它们对于映射局部纹理 (例如 photo2vangogh 和 photo2portrait) 的样式转换任务是成功的, 但通常对于野外图像中形状变化较大的图像翻译任务 (例如, selfie2anime 和 cat2dog) 而言是不成功的。因此, 通常需要通过限制数据分布的复杂度来避免图像分割和对齐等预处理步骤 (Huang 等人 (2018); Liu 等人 (2017))。此外, 现有方法如 DRIT (Lee 等人 (2018)) 无法

*大多数工作在 NCSoft 完成

†关联作者

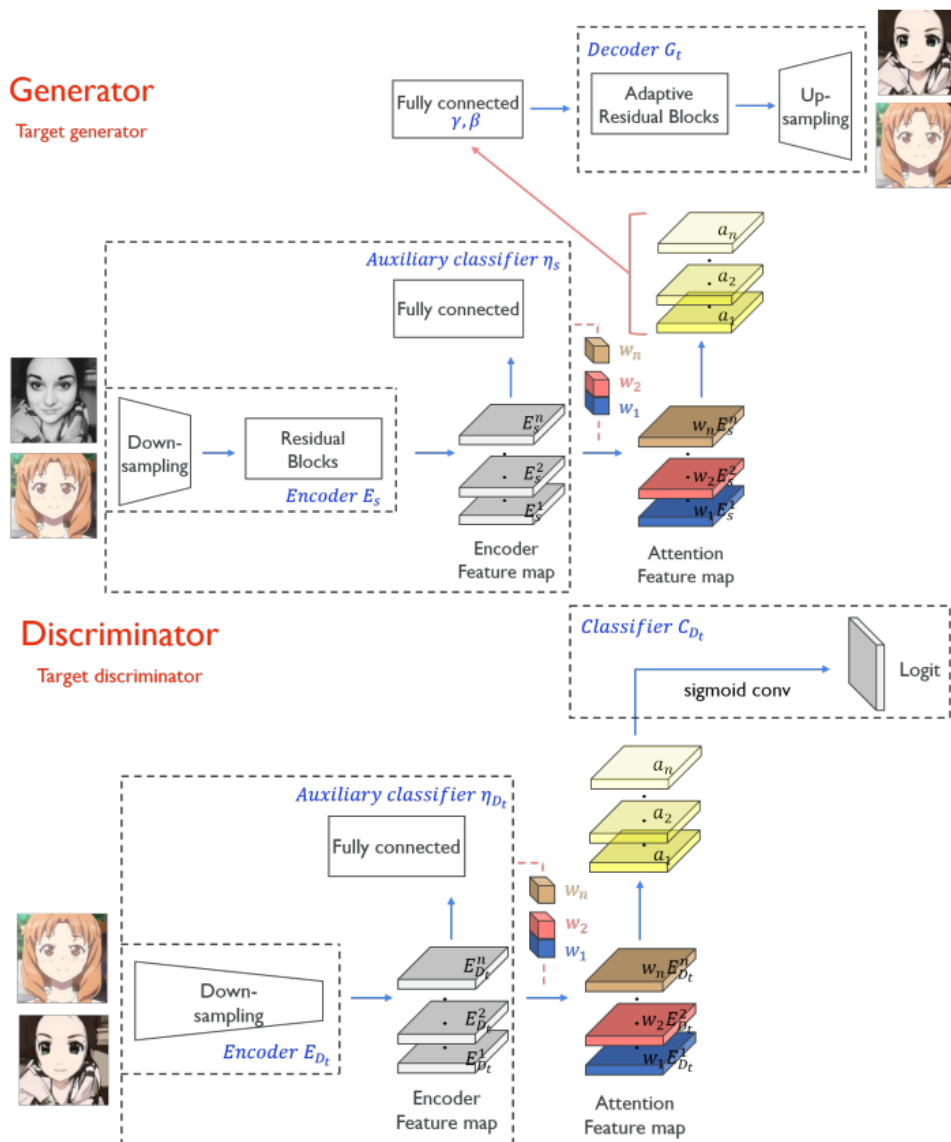


图 1: U-GAT-IT 的模型架构。详细的含义在模型架构章节中描述。

使用固定的网络架构和超参数，获得既可以保留形状的图像平移（例如，horse2zebra）也可以更改形状的图像平移（例如，cat2dog）的期望结果。它需要针对特定数据集调整网络结构或超参数设置。

在这项工作中，我们提出了一种无监督的图像到图像翻译的新方法，该方法以端到端的方式结合了新的注意力模块和新的可学习的归一化函数。我们的模型基于辅助分类器获得的注意力图，通过区分源域和目标域，指导翻译专注于更重要的区域，而忽略次要区域。这些注意力图被嵌入到生成器和判别器中，以专注于语义上重要的区域，从而促进了形状转换。生成器中的注意力图引起了对专门区分这两个域的区域关注，而判别器中的注意力图则通过关注目标域中真实图像和伪图像之间的差异来帮助进行微调。除了注意力机制之外，我们还发现，对于形状和纹理变化量不同的各种数据集，归一化函数的选择对转换结果的质量有重大影响。受批量实例归一化（BIN）（Nam & Kim (2018)）的启发，我们提出了自适应层实例归一化（AdaLIN），其参数是在训练期间从数据集中自适应学习选择实例归一化（IN）和层归一化（LN）之间的适当比率。AdaLIN 函数可帮助我们的注意力导向模型灵活地控制形状和纹理的变化量。结果，我们的模型无需修改模型架构或超参数，就可以执行图像翻译任务，不仅满足整体更改，而且也满足进行大的形状更改的需求。在实验中，我们证明了与现有的最新模型相比，该方法在样式转换和对象变形方面均具有优势。我们工作的主要贡献可归纳如下：

- 我们提出了一种新的方法，通过新的注意力模块和新的归一化函数 AdaLIN 进行无监督的图像到图像翻译。
- 我们的注意力模块通过基于辅助分类器获得的注意力图，通过在源域和目标域之间进行区分来帮助模型了解要在哪里进行集中转换。
- AdaLIN 函数可帮助我们的注意力导向模型灵活地控制形状和纹理的变化量，而无需修改模型架构或超参数。

2 带有自适应层实例化归一化的未监督的注意力生成网络

我们的目标是训练一个函数 $G_{s \rightarrow t}$ ，该函数仅使用从每个域提取的未配对样本将图像从源域 X_s 映射到目标域 X_t 。我们的框架由两个生成器 $G_{s \rightarrow t}$ 和 $G_{t \rightarrow s}$ 以及两个判别器 D_s 和 D_t 组成。我们将注意力模块集成到生成器和判别器中。判别器中的注意力模块引导生成器将注意力集中在生成逼真的图像至至关重要的区域上。生成器中的注意力模块将注意力集中到与其他域区分开的区域。在这里，我们仅解释 $G_{s \rightarrow t}$ 和 D_t (见图 1)，反之亦然。

2.1 模型

2.1.1 生成器

设 $x \in \{X_s, X_t\}$ 代表来自源域和目标域的样本。我们的翻译模型 $G_{s \rightarrow t}$ 由一个编码器 E_s ，一个解码器 G_t 和一个辅助分类器 η_s 组成，其中 $\eta_s(x)$ 表示 x 来自 X_s 的概率。设 $E_s^k(x)$ 为编码器的第 k 个激活图， $E_s^{kij}(x)$ 为 (i, j) 处的值。受 CAM (Zhou 等人 (2016)) 的启发，辅助分类器被训练为通过使用全局平均池化和全局最大池化，来学习源域的第 k 个特征图的权重 w_s^k ，即： $\eta_s(x) = \sigma(\sum_k w_s^k \sum_{ij} E_s^{kij}(x))$ 。通过利用 w_s^k ，我们可以计算一组特定领域的关注特征图，如 $a_s(x) = w_s * E_s(x) = \{w_s^k * E_s^k(x) | 1 \leq k \leq n\}$ ，其中 n 是编码的特征图的数量。然后，我们的转换模型 $G_{s \rightarrow t}$ 等于 $G_t(a_s(x))$ 。受最近在归一化层中使用仿射变换参数并结合归一化函数的工作的启发 (Huang & Belongie (2017); Nam & Kim (2018))，我们为残差块配备了参数为 AdaLIN 的残差块，并通过完全连接来动态计算注意力图上的特征。

$$AdaLIN(a, \gamma, \beta) = \gamma \cdot (\rho \cdot \hat{a}_I + (1 - \rho) \cdot \hat{a}_L) + \beta,$$

$$\hat{a}_I = \frac{a - \mu_I}{\sqrt{\sigma_I^2 + \epsilon}}, \hat{a}_L = \frac{a - \mu_L}{\sqrt{\sigma_L^2 + \epsilon}}, \quad (1)$$

$$\rho \leftarrow clip_{[0,1]}(\rho - \tau \Delta \rho)$$

其中 μ_I 、 μ_L 和 σ_I 、 σ_L 分别是按通道、按层的平均值和标准差， γ 和 β 是由全连接层生成的参数， τ 是学习率，并且 $\Delta \rho$ 表示由优化器确定的参数更新向量 (例如，梯度)。 ρ 只需在参数更新步骤中施加界限，即将值限制在 $[0, 1]$ 的范围内。生成器会调整该值，以便在 IN 非常重要的任务中，使其值接近 1，而在 LN 重要的任务中，其值接近 0。 ρ 的值在解码器的残差块中初始化为 1，在解码器的上采样块中初始化为 0。

将内容特征转移到样式特征上的最佳方法是应用“白化和着色变换” (Whitening and Coloring Transform, WCT) (Li 等人 (2017b))，但是由于协方差矩阵和矩阵的逆的计算，计算成本很高。尽管 AdaLN (Huang & Belongie (2017)) 比 WCT 快得多，但由于它假设特征通道之间不相关，因此它对 WCT 而言次优。因此，转移的特征包含更多的内容模式。另一方面，LN (Ba 等人 (2016)) 并未假设通道之间存在不相关性，但是

作为 ICLR 2020 的会议论文发表

有时它不能很好地保持原始域的内容结构,因为它仅考虑特征图的全局统计信息。为了克服这个问题,我们提出的归一化技术 AdaLIN 通过选择性地保留或更改内容信息来结合 AdaIN 和 LN 的优点,这有助于解决各种图像到图像的翻译问题。

2.1.2 判别器

令 $x \in \{X_t, G_{s \rightarrow t}(X_s)\}$ 表示来自目标域和转换后的源域的样本。与其他转换模型相似,判别器 D_t 是多尺度模型,由编码器 E_{D_t} , 分类器 C_{D_t} 和辅助分类器 η_{D_t} 组成。与其他转换模型不同, $\eta_{D_t}(x)$ 和 $D_t(x)$ 都经过训练以区分 x 来自 X_t 还是 $G_{s \rightarrow t}(X_s)$ 。给定样本 x , $D_t(x)$ 在通过 $\eta_{D_t}(x)$ 训练的编码特征图 $E_{D_t}(x)$ 上使用 w_{D_t} 获得注意力特征图 $a_{D_t}(x) = w_{D_t} * E_{D_t}(x)$ 。然后,我们的判别器 $D_t(x)$ 等于 $C_{D_t}(a_{D_t}(x))$ 。

2.2 损失函数

我们模型的全部目标包括四个损失函数。在这里,我们使用最小二乘 GAN (Mao 等人 (2017)) 来稳定训练,而不是使用原始 GAN 损失。

对抗损失 对抗损失用于使翻译图像的分布与目标图像分布匹配:

$$L_{lsGAN}^{s \rightarrow t} = (\mathbb{E}_{x \sim X_t} [(D_t(x))^2] + \mathbb{E}_{x \sim X_s} [(1 - D_t(G_{s \rightarrow t}(x)))^2]). \quad (2)$$

循环损失 为了减轻模式坍塌问题,我们将周期一致性约束应用于生成器。给定一个图像 $x \in X_s$, 在 x 从 X_s 到 X_t 以及从 X_t 到 X_s 的顺序转换之后,图像应成功转换回原始域:

$$L_{cycle}^{s \rightarrow t} = \mathbb{E}_{x \sim X_s} [|x - G_{t \rightarrow s}(G_{s \rightarrow t}(x))|_1]. \quad (3)$$

身份损失 为了确保输入图像和输出图像的颜色分布相似,我们将身份一致性约束应用于生成器。给定图像 $x \in X_t$, 使用 $G_{s \rightarrow t}$ 转换 x 后,图像不应更改:

$$L_{identity}^{s \rightarrow t} = \mathbb{E}_{x \sim X_t} [|x - G_{s \rightarrow t}(x)|_1]. \quad (4)$$

CAM 损失 通过利用来自辅助分类器 η_s 和 η_{D_t} 的信息,给定图像 $x \in \{X_s, X_t\}$ 。 $G_{s \rightarrow t}$ 和 D_t 知道他们需要改进的地方或在当前状态下两个域之间最大的区别是什么:

$$L_{cam}^{s \rightarrow t} = -(\mathbb{E}_{x \sim X_s} [\log(\eta_s(x))] + \mathbb{E}_{x \sim X_t} [\log(1 - \eta_s(x))]), \quad (5)$$

$$L_{cam}^{D_t} = \mathbb{E}_{x \sim X_t} [(\eta_{D_t}(x))^2] + \mathbb{E}_{x \sim X_s} [(1 - \eta_{D_t}(G_{s \rightarrow t}(x)))^2]. \quad (6)$$

完整目标 最后,我们联合训练编码器,解码器,判别器和辅助分类器,以优化最终目标:

$$\min_{G_{s \rightarrow t}, G_{t \rightarrow s}, \eta_s, \eta_t} \max_{D_s, D_t, \eta_{D_s}, \eta_{D_t}} \lambda_1 L_{lsGAN} + \lambda_2 L_{cycle} + \lambda_3 L_{identity} + \lambda_4 L_{cam}, \quad (7)$$

其中 $\lambda_1 = 1$; $\lambda_2 = 10$; $\lambda_3 = 10$; $\lambda_4 = 1000$ 。这里, $L_{lsGAN} = L_{lsGAN}^{s \rightarrow t} + L_{lsGAN}^{t \rightarrow s}$, 其他损失的定义方式也类似 (L_{cycle} , $L_{identity}$ 和 L_{cam})

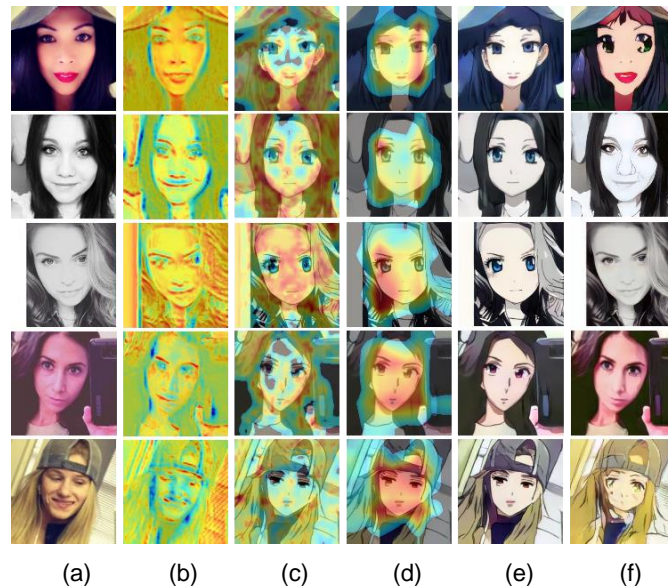


图 2: 消融实验中显示的注意力图及其效果的可视化: (a) 源图像, (b) 生成器的注意力图, (c-d) 判别器的局部和全局注意力图。 (e) 使用 CAM 的结果, (f) 不使用 CAM 的结果。

3 实验

3.1 基线模型

我们已将我们的方法与各种模型进行了比较, 包括 CycleGAN (Zhu 等人 (2017)), UNIT (Liu 等人 (2017)), MUNIT (Huang 等人 (2018)), DRIT (Lee 等人 (2018)), AGGAN (Mejjati 等人 (2018)) 和 CartoonGAN (Chen 等人 (2018))。所有基线方法都是使用作者的代码实现的。

3.2 数据集

我们用五个未配对的图像数据集 (包括四个代表性图像翻译数据集) 和一个新创建的由真实照片和动画作品 (即 selfie2anime) 组成的数据集评估了每种方法的性能。所有图像均调整为 256 x 256 进行训练。有关实验的每个数据集, 请参见附录 C。

3.3 实验结果

我们首先分析了所提出模型中注意力模块和 AdaLIN 的作用。然后, 我们将模型的性能与上一节中列出的其他无监督图像转换模型进行比较。为了评估翻译图像的视觉质量, 我们进行了一项用户研究。要求用户从五种不同方法生成的图像中选择最佳图像。补充材料中包含更多将我们的模型与其他模型进行比较的结果示例。

3.3.1 CAM 分析

首先, 我们进行消融研究, 以确认生成器和判别器中使用的注意力模块的益处。如图 2 (b) 所示, 注意力特征图帮助生成器将精力集中在与目标域更具区分性的源图像区域上, 例如眼睛和嘴巴。同时, 我们可以看到判别器集中注意力的区域, 通过可视化区域的局部和全局注意力图来确定目标图像是真是假,

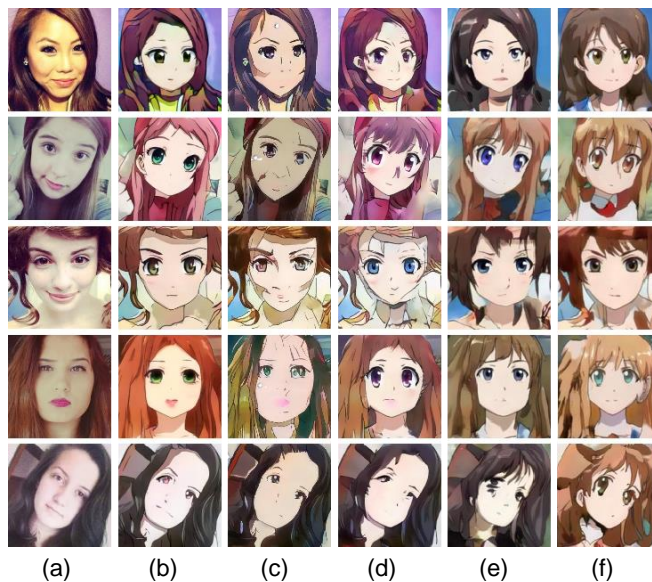


图 3: 使用每个归一化函数的结果比较: (a) 源图像, (b) 我们的结果, (c) 仅在带有 CAM 的解码器中使用 IN 的结果, (d) 仅在带有 CAM 的解码器中使用 LN 的结果, (e) 仅在带 CAM 的解码器中使用 AdaIN 的结果, (f) 仅在带 CAM 的解码器中使用 GN 的结果。

分别如图 2 (c) 和 (d) 所示。生成器可以使用这些注意力微调判别器关注的区域。请注意, 我们结合了来自两个具有不同感受野大小的判别器的全局和局部注意力图。这些注意力图可以帮助生成器捕获全局结构 (例如, 面部区域和眼睛附近) 以及局部区域。有了这些信息, 某些区域的翻译就会更加谨慎。图 2 (e) 所示的注意力模块的结果验证了在图像翻译任务中利用注意力特征图的有效效果。另一方面, 如果不使用如图 2 (f) 所示的注意力模块, 则可以看到眼睛未对准或完全不进行翻译。

3.3.2 AdaLIN 分析

如附录 B 所述, 我们仅将 AdaLIN 应用于生成器的解码器。残余块在解码器中的作用是嵌入特征, 而上采样卷积块在解码器中的作用是从嵌入特征生成目标域图像。如果阈值参数 ρ 的学习值更接近于 1, 则意味着相应的层相比 LN 更依赖于 IN。同样, 如果 ρ 的学习值更接近于 0, 则意味着相应的层相比 IN 更依赖于 LN。如图 3 (c) 所示, 在解码器中仅使用 IN 的情况下, 由于在残差中使用了按通道进行归一化的特征统计, 因此很好地保留了源域的特征块 (例如耳环和下颌骨周围的阴影)。然而, 由于全局风格不能被上采样卷积块的 IN 捕获, 所以转换为目标域风格的量在某种程度上是不足的。另一方面, 如图 3 (d) 所示, 如果我们在解码器中仅使用 LN, 则可以借助上采样卷积中使用的逐层归一化特征统计信息来充分传输目标域样式。但是, 通过在残差块中使用 LN 可以减少保留源域图像的特征。对两种极端情况的分析表明, 在特征表示层中更多地依赖 IN 而不是 LN 来保留源域的语义特征是有益的, 而实际上从特征生成图像的上采样层则相反嵌入。因此, 在无监督的图像到图像转换任务中, 我们提议的 AdaLIN 根据源域和目标域来调整解码器中 IN 和 LN 的比率更为理想。此外, 图 3 (e), (f) 分别是使用 AdaIN 和组归一化 (GN) 的结果 (Wu & He (2018)), 与这些方法相比, 我们的方法显示出更好的结果。

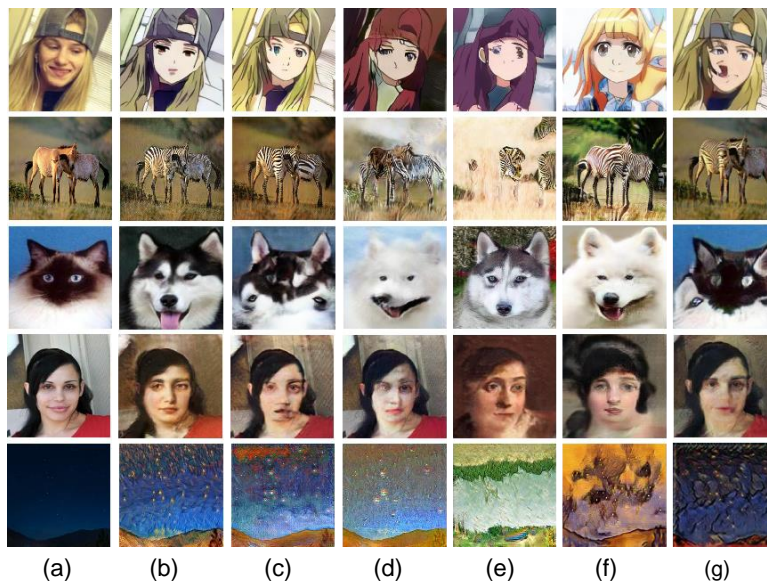


图 4: 对五个数据集的视觉比较。从上到下: selfie2anime, horse2zebra, cat2dog, photo2portrait 和 photo2vangogh。 (a) 源图像, (b) U-GAT-IT, (c) CycleGAN, (d) UNIT, (e) MUNIT, (f) DRIT, (g) AGGAN

表 1: 核初始距离 (*Kernel Inception Distance*) $\times 100 \pm \text{std.} \times 100$ 为了消融我们的模型。结果越低越好。有一些记号; GN: 组归一化, G_CAM: 生成器的 CAM, D_CAM: 判别器的 CAM。

Model	selfie2anime	anime2selfie
U-GAT-IT	11.61 \pm 0.57	11.52 \pm 0.57
U-GAT-IT w/ IN	13.64 \pm 0.76	13.58 \pm 0.8
U-GAT-IT w/ LN	12.39 \pm 0.61	13.17 \pm 0.8
U-GAT-IT w/ AdaIN	12.29 \pm 0.78	11.81 \pm 0.77
U-GAT-IT w/ GN	12.76 \pm 0.64	12.30 \pm 0.77
U-GAT-IT w/o CAM	12.85 \pm 0.82	14.06 \pm 0.75
U-GAT-IT w/o G_CAM	12.33 \pm 0.68	13.86 \pm 0.75
U-GAT-IT w/o D_CAM	12.49 \pm 0.74	13.33 \pm 0.89

此外, 如表 1 所示, 我们通过使用核起始距离 (KID) 进行的消融研究证明了 selfie2anime 数据集中注意力模块和 AdaLIN 的性能 (Binkowski 等人 (2018))。我们的模型实现了最低的 KID 值。即使将注意力模块和 AdaLIN 分开使用, 我们也可以看到我们的模型比其他模型表现更好。但是, 一起使用时, 性能会更好。

3.3.3 定性评估

为了进行定性评估, 我们还进行了感知研究。向 135 名参与者展示了来自不同方法的翻译结果, 包括带有源图像的拟议方法, 并要求他们选择最佳翻译图像到目标域。我们仅向参与者告知目标域的名称, 即动画, 狗和斑马。但是, 为肖像和 Van Gogh 数据集提供了一些目标域的示例图像, 作为确保正确判断的最小信息。表 2 表明, 除了 photo2vangogh 以外, 该方法的得分显著提高, 但在人类感知研究中与其他方法相当。在图 4 中, 我们展示了每种方法的图像翻译结果, 以进行性能比较。U-GAT-IT 可以通过更多地关注源之间的不同区域来生成未失真的图像

表 2: 用户研究对翻译图像的偏好得分。

Model	selfie2anime	horse2zebra	cat2dog	photo2portrait	photo2vangogh
U-GAT-IT	73.15	73.56	58.22	30.59	48.96
CycleGAN	20.07	23.07	6.19	26.59	27.33
UNIT	1.48	0.85	18.63	32.11	11.93
MUNIT	3.41	1.04	14.48	8.22	2.07
DRIT	1.89	1.48	2.48	2.48	9.70

表 3: 核初始距离 $\times 100 \pm \text{std.} \times 100$ 为了差异化图像转换模式。结果越低越好。

Model	selfie2anime	horse2zebra	cat2dog	photo2portrait	photo2vangogh
U-GAT-IT	11.61 \pm 0.57	7.06 \pm 0.8	7.07 \pm 0.65	1.79 \pm 0.34	4.28 \pm 0.33
CycleGAN	13.08 \pm 0.49	8.05 \pm 0.72	8.92 \pm 0.69	1.84 \pm 0.34	5.46 \pm 0.33
UNIT	14.71 \pm 0.59	10.44 \pm 0.67	8.15 \pm 0.48	1.20 \pm 0.31	4.26 \pm 0.29
MUNIT	13.85 \pm 0.41	11.41 \pm 0.83	10.13 \pm 0.27	4.75 \pm 0.52	13.08 \pm 0.34
DRIT	15.08 \pm 0.62	9.79 \pm 0.62	10.92 \pm 0.33	5.85 \pm 0.54	12.65 \pm 0.35
AGGAN	14.63 \pm 0.55	7.58 \pm 0.71	9.84 \pm 0.79	2.33 \pm 0.36	6.95 \pm 0.33
CartoonGAN	15.85 \pm 0.69	-	-	-	-

Model	anime2selfie	zebra2horse	dog2cat	portrait2photo	vangogh2photo
U-GAT-IT	11.52 \pm 0.57	7.47 \pm 0.71	8.15 \pm 0.66	1.69 \pm 0.53	5.61 \pm 0.32
CycleGAN	11.84 \pm 0.74	8.0 \pm 0.66	9.94 \pm 0.36	1.82 \pm 0.36	4.68 \pm 0.36
UNIT	26.32 \pm 0.92	14.93 \pm 0.75	9.81 \pm 0.34	1.42 \pm 0.24	9.72 \pm 0.33
MUNIT	13.94 \pm 0.72	16.47 \pm 1.04	10.39 \pm 0.25	3.30 \pm 0.47	9.53 \pm 0.35
DRIT	14.85 \pm 0.60	10.98 \pm 0.55	10.86 \pm 0.24	4.76 \pm 0.72	7.72 \pm 0.34
AGGAN	12.72 \pm 1.03	8.80 \pm 0.66	9.45 \pm 0.64	2.19 \pm 0.40	5.85 \pm 0.31

通过利用注意力模块来确定目标域。请注意, 来自 CycleGAN 的结果表明, 两只斑马或狗的眼睛周围的区域变形。此外, 使用 U-GAT-IT 的翻译结果在视觉上优于其他方法, 同时保留了源域的语义特征。值得注意的是, MUNIT 和 DRIT 的结果与源图像非常不同, 因为它们生成的图像带有用于多样性的随机样式代码。此外, 应该强调的是, U-GAT-IT 对五个不同的数据集都应用了相同的网络体系结构和超参数, 而其他算法则使用预设的网络或超参数进行训练。通过用户研究的结果, 我们表明注意力模块和 AdaLIN 的组合使我们的模型更加灵活。

3.3.4 定量评估

为了进行定量评估, 我们使用了最近提出的 KID, 它可以计算真实图像和生成图像的特征表示之间的最大均方差的平方。特征表示是从 Inception 网络中提取的 (Szegedy 等人 (2016))。与 Frechet 起始距离 (Heusel 等人, 2017) 相比, KID 具有无偏估计量, 这使其更可靠, 尤其是在测试图像少于起始特征维数的情况下。较低的 KID 表示真实图像和生成的图像之间共享的视觉相似度更高 (Mejjati 等人 (2018))。因此, 如果翻译正确, 则 KID 在多个数据集中将具有较小的值。表 3 显示, 除了诸如 photo2vangogh 和 photo2portrait 之类的样式转换任务外, 该方法的 KID 得分最低。但是, 与最低分数没有太大区别。此外, 与 UNIT 和 MUNIT 不同, 我们可以看到源 \rightarrow 目标, 目标 \rightarrow 源代码翻译都是稳定的。与最近基于注意力的方法 AGGAN 相比, U-GAT-IT 的 KID 甚至更低。与 U-GAT-IT 不同, AGGAN 对于形状变化 (如 dog2cat 和 anime2selfie) 的变换产生的性能不佳, 而 U-GAT-IT 的注意力模块的重点是不区分背景和前景, 而是区分

两个域。如补充材料中所示, CartoonGAN 仅将图像的整体颜色更改为动画样式, 但是与自拍照相比, 动画的最大特点是眼睛根本没有变化。因此, CartoonGAN 具有较高的 KID。

4 结论

在本文中, 我们提出了一种带有注意力模块和 AdaLIN 的无监督图像到图像转换 (U-GAT-IT), 它可以在具有固定网络架构和超参数的各种数据集中产生更直观的视觉效果。对各种实验结果的详细分析支持我们的假设, 即由辅助分类器获得的注意力图可以指导生成器将更多的注意力集中在源域和目标域之间的不同区域。此外, 我们发现自适应层实例归一化 (AdaLIN) 对于转换包含不同数量的几何形状和样式变换的各种数据集至关重要。通过实验, 我们表明, 相比于现有的基于 GAN 的最新模型, 该方法在无监督的图像到图像翻译任务中具有优越性。

参考文献

- Asha Anoopshah, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 783–790, 2018.
- Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In International Conference on Machine Learning, pp. 214–223, 2017.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.
- David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717, 2017.
- Mikołaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In International Conference on Learning Representations, 2018. URL <https://openreview.net/forum?id=r1IUozWCW>.
- Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9465–9474, 2018.
- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Star-gan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797, 2018.
- Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2): 295–307, 2016.
- Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In International Conference on Learning Representations, 2017. URL <https://openreview.net/forum?id=BJO-BuT1g>.
- Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423, 2016.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2672–2680, 2014.

- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.
- Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.
- Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pp. 172–189, 2018.
- Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107, 2017.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=Hk99zCeAb>.
- Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.
- Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1857–1865, 2017. URL <http://proceedings.mlr.press/v70/kim17a.html>.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL <http://arxiv.org/abs/1412.6980>.
- Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for auto-matic colorization. In *Proceedings of the European Conference on Computer Vision*, pp. 577–593. Springer, 2016.
- Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision*, pp. 35–51, 2018.
- Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, pp. 5501–5509, 2017a.
- Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, pp. 386–396, 2017b.
- Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pp. 700–708, 2017.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.
- Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2813–2821. IEEE, 2017.

- Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pp. 3697–3707, 2018.
- Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=B1QRgziT->.
- Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *Advances in Neural Information Processing Systems*, pp. 2563–2572, 2018.
- Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, 2014.
- Amelie´ Royer, Konstantinos Bousmalis, Stephan Gouws, Fred Bertsch, Inbar Moressi, Forrester Cole, and Kevin Murphy. Xgan: Unsupervised image-to-image translation for many-to-many mappings. arXiv preprint arXiv:1711.05139, 2017.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.
- Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Sk2Im59ex>.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.
- Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, September 2018.
- Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2849–2857, 2017.
- Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, pp. 649–666. Springer, 2016.
- Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics*, 36(4):119:1–119:11, 2017. doi: 10.1145/3072959.3073703. URL <https://doi.org/10.1145/3072959.3073703>.
- Junbo Jake Zhao, Michael” Mathieu, and Yann LeCun. Energy-based generative adversarial net-works. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=ryh9pmcee>.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929. IEEE, 2016.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

作为 ICLR 2020 的会议论文发表

A 相关工作

A.1 生成对抗网络

生成对抗网络 (GAN) (Goodfellow 等人 (2014)) 在各种图像生成方面都取得了令人印象深刻的结果 (Arjovsky 等人 (2017); Berthelot 等人 (2017); Karras 等人 (2018)); Zhao 等 (2017)), 譬如图像修复 (Iizuka 等 (2017)), 图像翻译 (Choi 等 (2018); Huang 等 (2018); Isola 等 (2017); Liu 等人 (2017); Wang 等人 (2018); Zhu 等人 (2017)) 任务等。在训练中, 生成器旨在生成逼真的图像, 以欺骗判别器, 而判别器则试图将生成的图像与真实图像区分开。各种多阶段生成模型 (Karras 等 (2018); Wang 等 (2018)) 和更好的训练目标 (Arjovsky 等 (2017); Berthelot 等 (2017); Mao 等 (2017)); Zhao 等人 (2017)) 已提出以生成更逼真的图像。在本文中, 在没有配对训练数据的情况下, 我们的模型使用 GAN 来学习从源域到明显不同的目标域的转变。

A.2 图像到图像转换

Isola 等人 (Isola 等人 (2017)) 提出了基于条件 GAN 的图像到图像翻译的统一框架。Wang 等人 (Wang 等人 (2018)) 提出了高分辨率版本的 pix2pix (Huang 等人 (2018); Kim 等人 (2017); Liu 等人 (2017); Taigman 等人 (2017); Zhu 等人 (2017)) 从未配对的数据集中学习图像翻译。CycleGAN (Zhu 等人 (2017)) 首次提出了循环一致性损失, 以强制进行一对一映射。UNIT (Liu et al. (2017)) 假设使用一个共享的潜在空间来处理无监督的图像翻译。但是, 仅当两个域具有相似的模式时, 此方法才能很好地执行。MUNIT (Huang 等人 (2018)) 通过将图像分解为领域不变的内容代码和捕获领域特定属性的样式代码, 可以扩展到多对多映射。MUNIT 合成分离的内容和样式以生成最终图像, 其中通过使用自适应实例规范化来改善图像质量 (Huang & Belongie (2017))。出于与 MUNIT 相同的目的, DRIT (Lee 等人 (2018)) 将图像分解为内容和样式, 因此可以进行多对多映射。唯一的区别是, 使用权重共享和内容区分器 (辅助分类器) 在两个域之间共享内容空间。尽管如此, 这些方法的性能 (Huang 等人 (2018); Liu 等人 (2017); Lee 等人 (2018)) 仅限于包含源域和目标域之间对齐图像良好的数据集。此外, AGGAN (Mejjati 等人 (2018)) 通过使用注意力机制来区分前景和背景, 从而改善了图像翻译的性能。但是, AGGAN 中的注意力模块无法帮助您改变图像中对象的形状。虽然 CartoonGAN (Chen 等人 (2018)) 在动画样式转换方面表现出良好的性能, 但它仅更改图像中线条的颜色, 色调和粗细。因此, 它不适合图像中的形状变化。

A.3 类激活图

Zhou 等人 (Zhou et al (2016)) 提出了使用 CNN 中的全局平均池化的类激活图 (CAM)。特定类别的 CAM 通过 CNN 显示判别图像区域, 以确定该类别。在这项工作中, 我们的模型通过使用 CAM 方法来区分两个域, 从而导致集约化图像区域的密集变化。但是, 我们不仅使用了全局平均池化, 而且还使用了全局最大池化来使结果更好。

A.4 正则化

最近的神经样式转换研究表明, CNN 特征统计数据 (例如 Gram 矩阵 (Gatys 等人, 2016)), 均值和方差 (Huang & Belongie, 2017) 可以用作图像样式的直接描述符。实例归一化 (IN) 具有通过直接归一化图像的特征统计来消除样式变化的作用, 并且在样式转换中比批归一化 (BN) 或层归一化 (LN) 的使用频率更高。最近的研究使用自适应实例归一化 (AdaIN) (Huang 和 Belongie (2017)), 条件实例归一化 (CIN) (Dumoulin 等人 (2017)) 和批处理实例归一化 (BIN) (Nam & Kim (2018)), 而不是单独使用 IN。在我们的工作中, 我们提出了一种自适应

层实例归一化 (AdaLIN) 函数可自适应地选择 IN 和 LN 之间的适当比率。通过 AdaLIN, 我们的注意力导向模型可以灵活地控制形状和纹理的变化量。

B 实现细节

B.1 网络架构

U-GAT-IT 的网络体系结构如表 4、5 和 6 所示。生成器的编码器由两个卷积层组成, 步幅大小为 2, 用于下采样和四个残差块。生成器的解码器由四个残差块和两个上采样卷积层组成, 步幅为 1。请注意, 我们分别对编码器使用实例归一化, 对解码器使用 AdaLIN。通常, 在分类问题中, LN 的性能不比批量归一化好 (Wu & He (2018))。由于辅助分类器是从生成器中的编码器连接的, 因此, 为了提高辅助分类器的准确性, 我们使用实例归一化 (最小批量大小为 1 的批量归一化) 代替 AdaLIN。谱归一化 (Miyato 等人 (2018)) 用于判别器。我们使用两种不同比例的 PatchGAN (Isola 等人 (2017)) 来区分网络, 该网络对本地 (70 x 70) 和全局 (286 x 286) 图像补丁是真实的还是假的进行分类。对于激活函数, 我们在生成器中使用 ReLU, 在判别器中使用 0.2 斜率的 Leaky ReLU。

B.2 训练

所有模型均使用 Adam (Kingma & Ba (2015)) 进行训练, 其中 $\beta_1 = 0.5$ 和 $\beta_2 = 0.999$ 。对于数据增强, 我们以 0.5 的概率水平翻转图像, 将其调整为 286 x 286, 然后随机裁剪为 256 x 256。所有实验的批次大小均设置为 1。我们以 0.0001 的固定学习率训练所有模型, 直到进行 500,000 次迭代, 然后线性衰减直至 1,000,000 次迭代。我们还使用权重衰减为 0.0001 的比率。权重按零中心正态分布初始化, 标准偏差为 0.02。

C 数据集细节

selfie2anime 自拍照数据集包含 46,836 张带有 36 个不同属性的自拍照图像。我们仅将女性的照片用作训练数据和测试数据。训练数据集的大小为 3400, 测试数据集的大小为 100, 图像大小为 256 x 256。对于动漫数据集, 我们首先从 Anime-Planet (见底部 1) 中检索了 69,926 个动画角色图像。在这些图像中, 使用动漫人脸检测器 (见底部 2) 提取了 27,023 张人脸图像。在仅选择女性角色图像并手动去除单色图像之后, 我们收集了两个女性动漫人脸图像数据集, 分别用于训练和测试的数据大小分别为 3400 和 100, 这与自拍照数据集的编号相同。最后, 通过应用基于 CNN 的图像超分辨率算法 (见底部 3), 将所有动漫人脸图像的大小调整为 256 x 256。

horse2zebra 和 **photo2vangogh** 这些数据集用于 CycleGAN (Zhu 等人 (2017))。每个类别的训练数据集大小: 1,067 (马), 1,334 (斑马), 6,287 (照片) 和 400 (梵高)。测试数据集包括 120 (马), 140 (斑马), 751 (照片) 和 400 (梵高)。注意, 梵高类的训练数据和测试数据是相同的。

cat2dog 和 **photo2portrait** 这些数据集用于 DRIT (Lee 等人 (2018))。每个类别的数据数量为 871 (猫), 1,364 (斑马), 6,452 (照片) 和 1,811 (梵高)。我们分别使用 120 (马), 140 (斑马), 751 (照片) 和 400 (梵高) 随机选择的图像作为测试数据。

¹<http://www.anime-planet.com/>

²https://github.com/nagadomi/lbpcascade_animeface

³<https://github.com/nagadomi/waifu2x>

D 补充实验结果

除了本文提供的结果外, 我们还在图 5、6、7、8、9、10、11 和 12 中显示了五个数据集的补充生成结果。

表 4: 生成器架构的详细信息。

Part	Input \rightarrow Output Shape	Layer Information
Encoder Down-sampling	$(h, w, 3) \rightarrow (h, w, 64)$	CONV-(N64, K7, S1, P3), IN, ReLU
	$(h, w, 64) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	CONV-(N128, K3, S2, P1), IN, ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	CONV-(N256, K3, S2, P1), IN, ReLU
Encoder Bottleneck	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	ResBlock-(N256, K3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	ResBlock-(N256, K3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	ResBlock-(N256, K3, S1, P1), IN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	ResBlock-(N256, K3, S1, P1), IN, ReLU
CAM of Generator	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 512)$	Global Average & Max Pooling, MLP-(N1), Multiply the weights of MLP
	$(\frac{h}{4}, \frac{w}{4}, 512) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	CONV-(N256, K1, S1), ReLU
γ, β	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (1, 1, 256)$	MLP-(N256), ReLU
	$(1, 1, 256) \rightarrow (1, 1, 256)$	MLP-(N256), ReLU
	$(1, 1, 256) \rightarrow (1, 1, 256)$	MLP-(N256), ReLU
Decoder Bottleneck	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	AdaResBlock-(N256, K3, S1, P1), AdaILN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	AdaResBlock-(N256, K3, S1, P1), AdaILN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	AdaResBlock-(N256, K3, S1, P1), AdaILN, ReLU
	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{4}, \frac{w}{4}, 256)$	AdaResBlock-(N256, K3, S1, P1), AdaILN, ReLU
Decoder Up-sampling	$(\frac{h}{4}, \frac{w}{4}, 256) \rightarrow (\frac{h}{2}, \frac{w}{2}, 128)$	Up-CONV-(N128, K3, S1, P1), LIN, ReLU
	$(\frac{h}{2}, \frac{w}{2}, 128) \rightarrow (h, w, 64)$	Up-CONV-(N64, K3, S1, P1), LIN, ReLU
	$(h, w, 64) \rightarrow (h, w, 3)$	CONV-(N3, K7, S1, P3), Tanh

表 5: 局部判别器的详细信息

Part	Input \rightarrow Output Shape	Layer Information
Encoder Down-sampling	$(h, w, 3) \rightarrow (\frac{h}{2}, \frac{w}{2}, 64)$	CONV-(N64, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{2}, \frac{w}{2}, 64) \rightarrow (\frac{h}{4}, \frac{w}{4}, 128)$	CONV-(N128, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{4}, \frac{w}{4}, 128) \rightarrow (\frac{h}{8}, \frac{w}{8}, 256)$	CONV-(N256, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{8}, \frac{w}{8}, 256) \rightarrow (\frac{h}{8}, \frac{w}{8}, 512)$	CONV-(N512, K4, S1, P1), SN, Leaky-ReLU
CAM of Discriminator	$(\frac{h}{8}, \frac{w}{8}, 512) \rightarrow (\frac{h}{8}, \frac{w}{8}, 1024)$	Global Average & Max Pooling, MLP-(N1), Multiply the weights of MLP
	$(\frac{h}{8}, \frac{w}{8}, 1024) \rightarrow (\frac{h}{8}, \frac{w}{8}, 512)$	CONV-(N512, K1, S1), Leaky-ReLU
Classifier	$(\frac{h}{8}, \frac{w}{8}, 512) \rightarrow (\frac{h}{8}, \frac{w}{8}, 1)$	CONV-(N1, K4, S1, P1), SN

表 6: 全局判别器的详细信息

Part	Input \rightarrow Output Shape	Layer Information
Encoder Down-sampling	$(h, w, 3) \rightarrow (\frac{h}{2}, \frac{w}{2}, 64)$	CONV-(N64, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{2}, \frac{w}{2}, 64) \rightarrow (\frac{h}{4}, \frac{w}{4}, 128)$	CONV-(N128, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{4}, \frac{w}{4}, 128) \rightarrow (\frac{h}{8}, \frac{w}{8}, 256)$	CONV-(N256, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{8}, \frac{w}{8}, 256) \rightarrow (\frac{h}{16}, \frac{w}{16}, 512)$	CONV-(N512, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{16}, \frac{w}{16}, 512) \rightarrow (\frac{h}{32}, \frac{w}{32}, 1024)$	CONV-(N1024, K4, S2, P1), SN, Leaky-ReLU
	$(\frac{h}{32}, \frac{w}{32}, 1024) \rightarrow (\frac{h}{32}, \frac{w}{32}, 2048)$	CONV-(N2048, K4, S1, P1), SN, Leaky-ReLU
CAM of Discriminator	$(\frac{h}{32}, \frac{w}{32}, 2048) \rightarrow (\frac{h}{32}, \frac{w}{32}, 4096)$	Global Average & Max Pooling, MLP-(N1), Multiply the weights of MLP
	$(\frac{h}{32}, \frac{w}{32}, 4096) \rightarrow (\frac{h}{32}, \frac{w}{32}, 2048)$	CONV-(N2048, K1, S1), Leaky-ReLU
Classifier	$(\frac{h}{32}, \frac{w}{32}, 2048) \rightarrow (\frac{h}{32}, \frac{w}{32}, 1)$	CONV-(N1, K4, S1, P1), SN

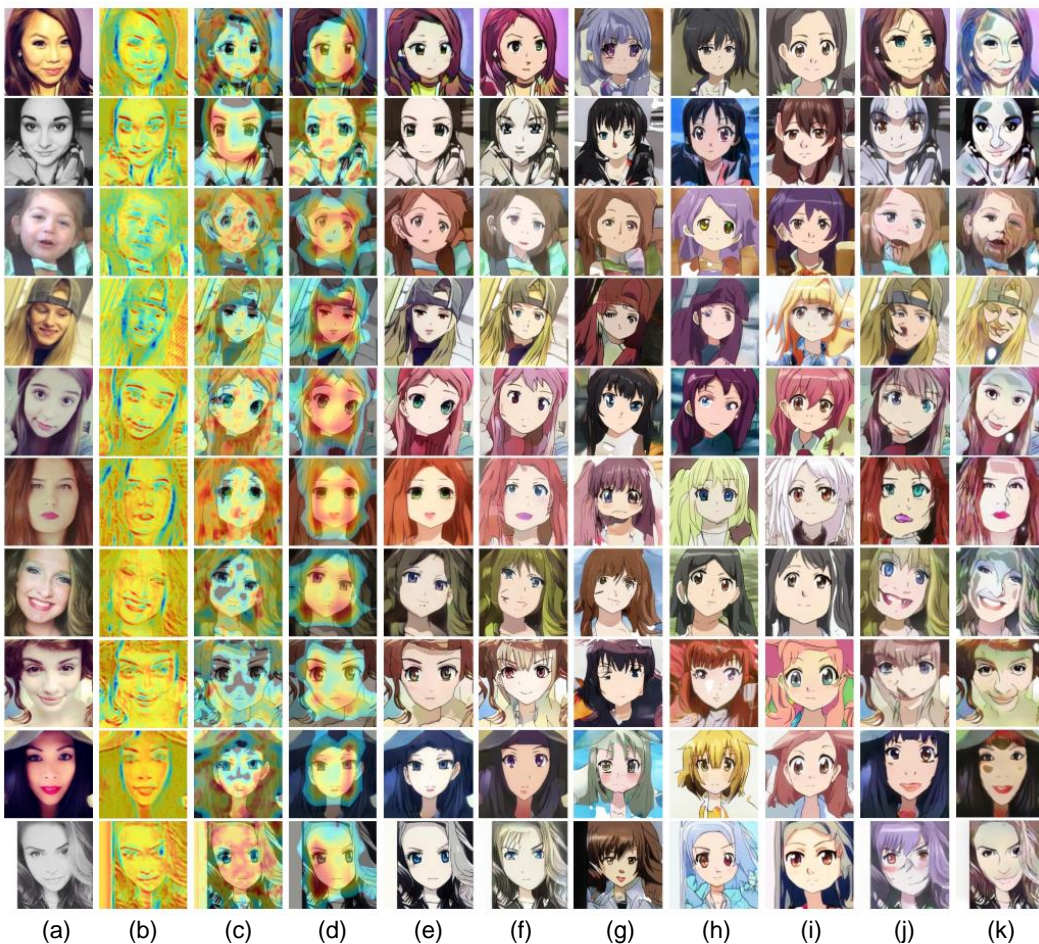


图 5: 带有注意特征图的 selfie2anime 的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018)), (k) CartoonGAN (Chen 等人 (2018))。

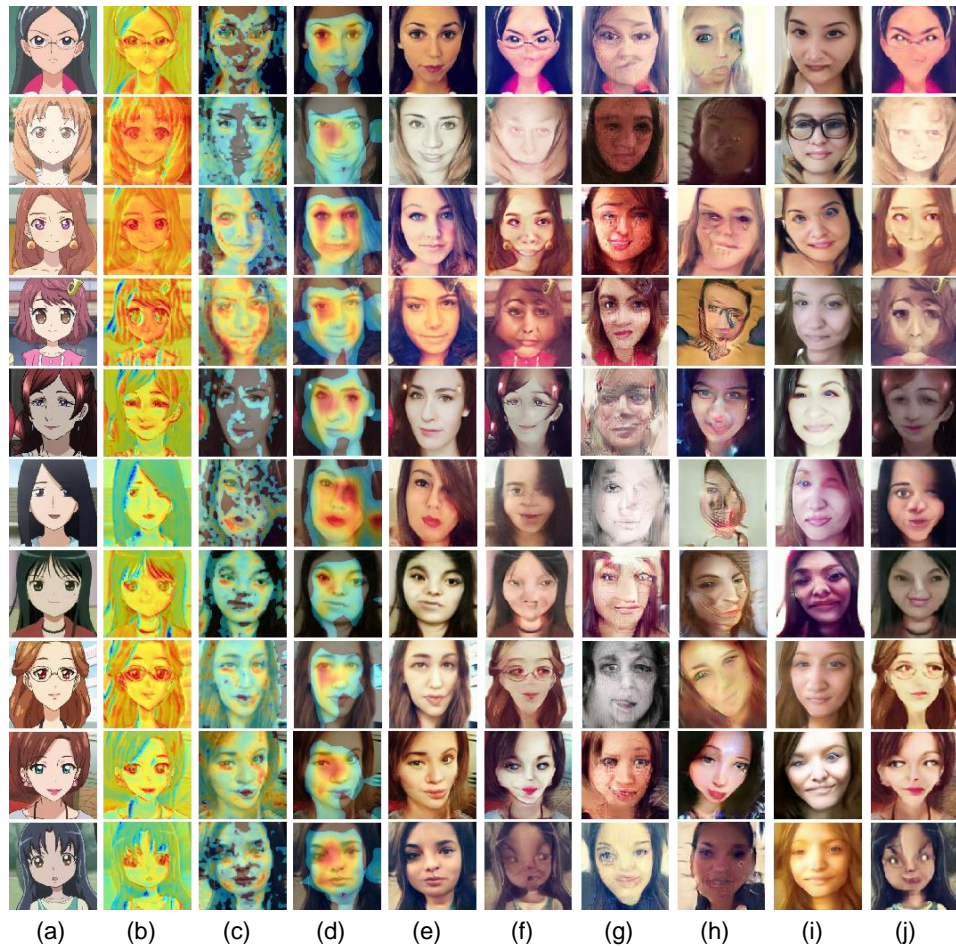


图 6: anime2selfie 与注意力特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018))。

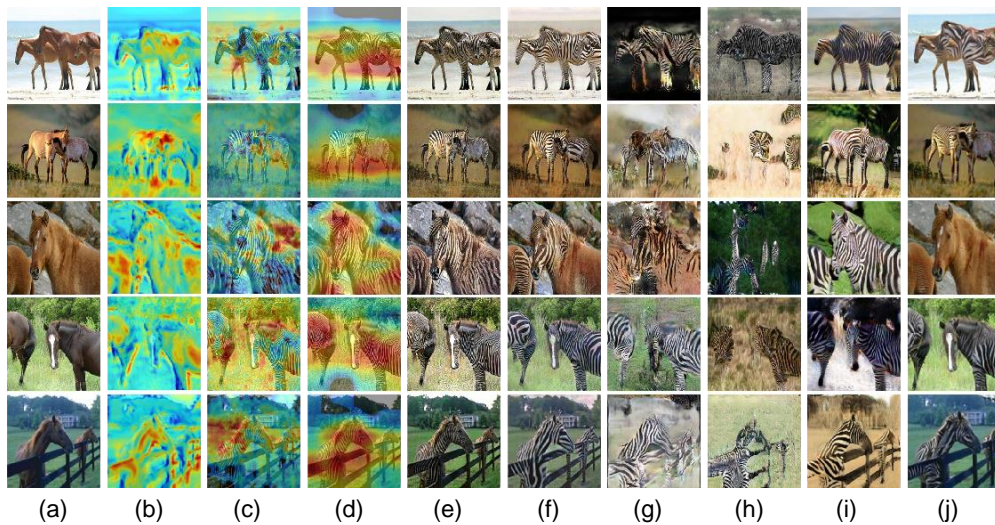


图 7: horse2zebra 与注意力特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018))))。

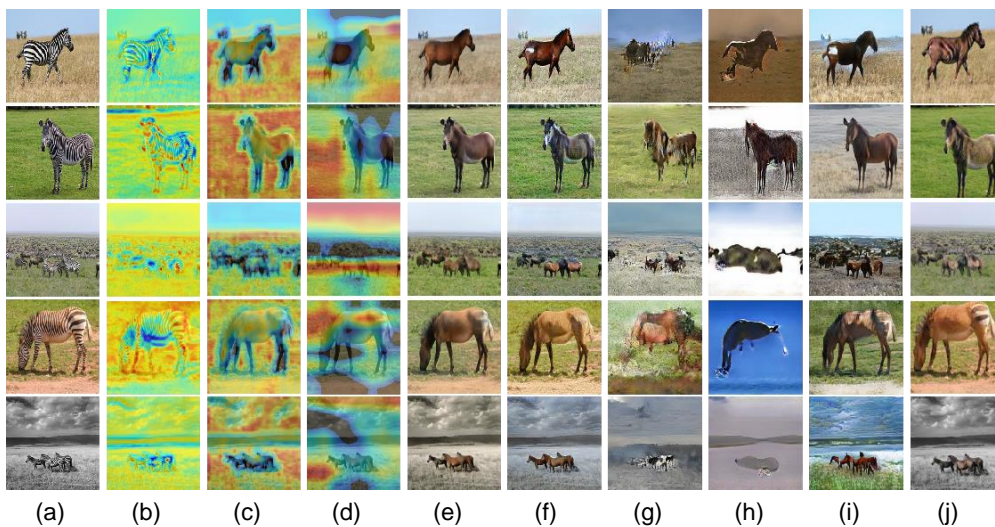


图 8: 斑马线 with 注意特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018))))。

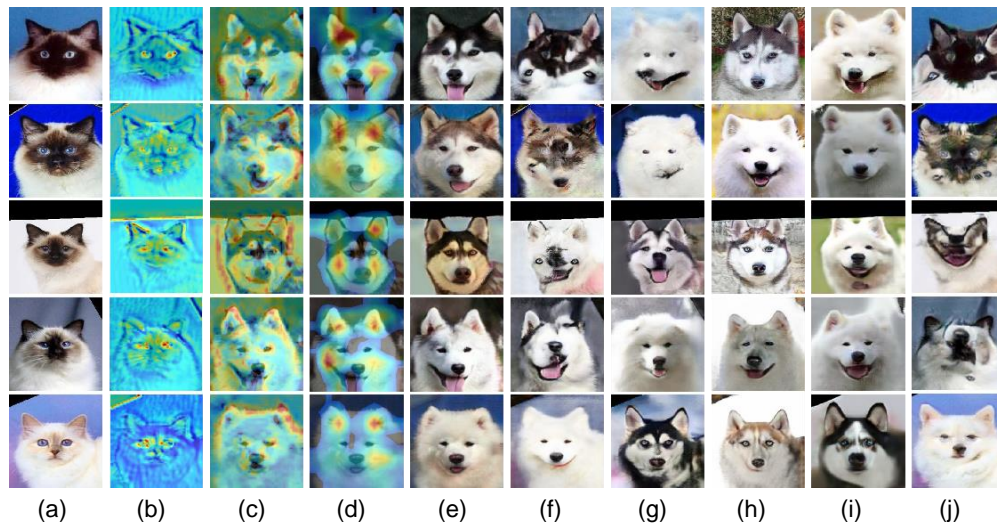


图 9: cat2dog 与注意力特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018)))。

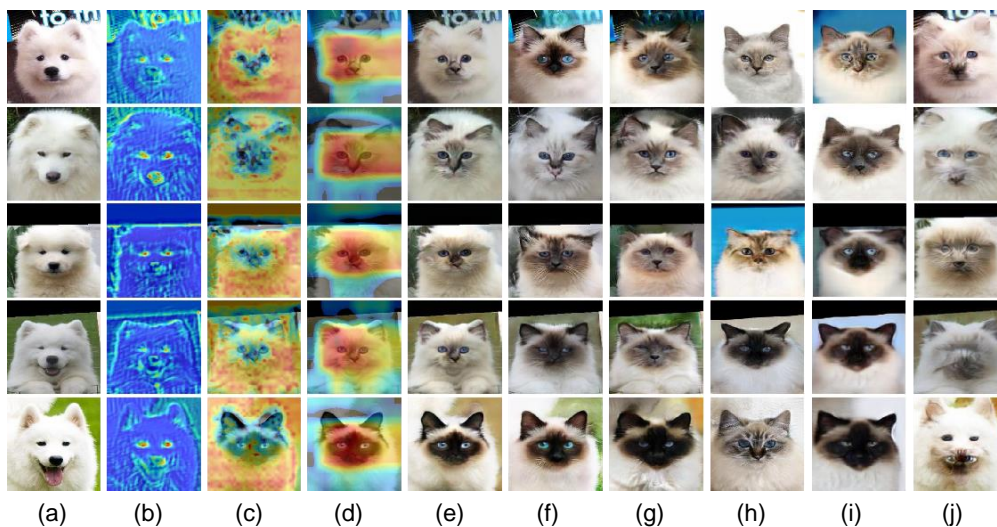


图 10: dog2cat 与注意力特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018)))。

作为 ICLR 2020 的会议论文发表

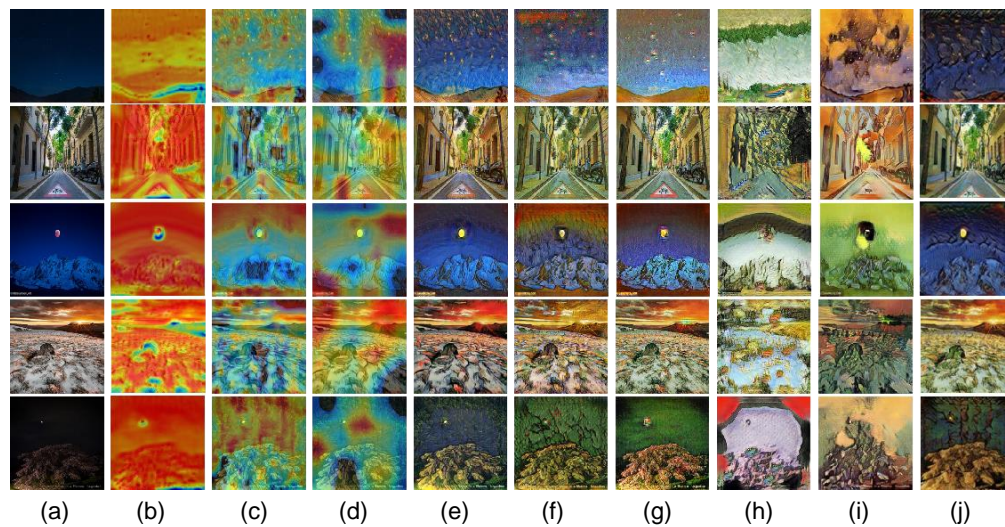


图 11: photo2vangogh 与注意力特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)) (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018))。

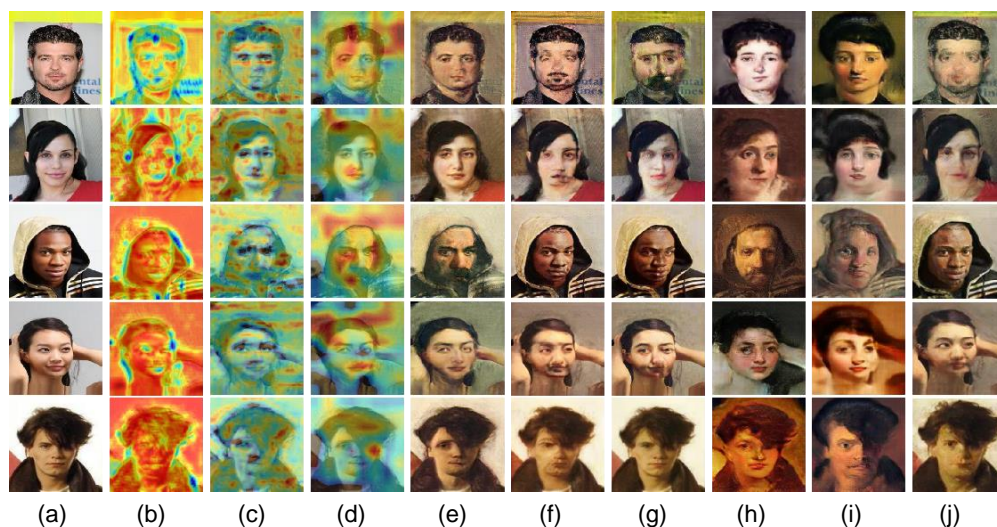


图 12: photo2portrait 与注意力特征图的视觉比较。(a) 源图像, (b) 生成器的注意力图, (cd) 判别器的局部和全局注意力图, (e) 我们的结果, (f) CycleGAN (Zhu 等人 (2017)), (g) UNIT (Liu 等人 (2017)), (h) MUNIT (Huang 等人 (2018)), (i) DRIT (Lee 等人 (2018)), (j) AGGAN (Mejjati 等人 (2018))。