基于 DINO-ViT 的语义风格迁移

摘要

视觉外观转移指的是将特定图像的视觉外观转移到另一个与其语义相关的图像上，同时保留该图像的原有结构。在两个图像之间存在明显的姿势、外观和形状差异的情况下，语义信息的提取至关重要。本文对 Tumanyan 等人 [14] 发表在 CVPR 2022 的 Splice-ViT 进行解读，并通过方法复现来充分认识 Vision Transformers (ViTs) 的潜在语义及其对结构与外观特征的解耦表达能力。首先介绍一种基于自监督方式训练的 ViT 并说明该 ViT 相比其他视觉模型的优越性；接着阐述在这种自监督 ViT 特征空间下的结构与外观表示；然后阐述基于解耦的结构与外观特征实现的视觉外观转移方法 Splice。最后通过方法复现以及大量实验结果充分地体现 Splice 方法的风格迁移能力，并讨论该方法存在的缺陷以及可能的改进与泛化策略。

关键词：图像生成；风格迁移；Vision Transformers；深度特征

1 引言

随着深度学习的快速发展，风格迁移领域不断涌现新的视觉任务和创新方法。其中，视觉外观转移作为一个关键任务，旨在将一个特定图像的视觉外观转移到与其语义相关的另一个图像上，同时保留原始图像的结构特征。这项任务在处理具有显著姿势、外观和形状差异的两个图像时，对语义信息的准确提取至关重要。为了解决这一挑战，Tumanyan 等人 [14] 提出了 Splice 方法，并将其最新成果发表在 CVPR 2022 上。该方法采用了一种基于自监督学习的 Vision Transformers (ViTs [3])，即 DINO [2]，有效地提取图像的视觉外观特征和语义结构特征。通过设计简单而强大的损失函数成功训练生成器，使其能够生成有目标图像的视觉外观的同时，保留输入图像的原有结构的结果图像。

本文提炼并总结 Splice 方法，根据该方法的特征选择思路解释 DINO 学习的潜在语义及其对结构与外观特征的解耦表达能力，然后阐述基于解耦的结构与外观特征实现的视觉外观转移方法。本文对该迁移方法进行复现，并展示大量视觉外观转移方法在不同语义图像上的迁移效果。

本文接下来在第2章中简述近几年基于深度学习的风格迁移以及有关 ViTs 特征表达的相关工作；接着在第3章分析 ViTs 对结构与外观表达及其解耦能力；然后在第4章展开 Splice 迁移方法的介绍；最后在第5章展示结果并对 Splice 方法存在的缺陷以及可能的改进与泛化策略展开讨论。

2 相关工作

2.1 风格迁移

图像的风格迁移主要指将一幅图像的艺术风格转移到另一个图像上，其最早可以追溯到图像类比 [7]。Hertzmann 等人 [7] 提出的图像类比方法中，旨在参考一对图像中原始图像到目标图像的风格变化，并将这种变化模式迁移到给定的图像，使目标图像的风格与目标图像的风格类似。随着深度学习的发展，Gatys 等人 [5] 首次提出使用深度特征构建风格指导，实现惊人的风格化效果。这种风格迁移方法的任务是给定一个内容图像和一个风格图像并将风格图像的艺术风格迁移到内容图像上，同时保留内容图像的结构。Gatys 等人 [5] 的方法中使用预训练的 VGG [12] 特征表示风格并利用其自相似性来捕捉结构，通过优化的方式来实现艺术风格的转移。与从类似，Kolkin 等人 [10] 提出了 STROTSS 方法，改进了这种风格迁移方法的效果。另外，Kim 等人 [9] 关注内容的变形提出 DST 方法。然而基于优化的方法效率低，为此 Huang 等人 [8] 提出了 AdaIN 方法，通过特征调制的方式进行风格化，并训练了一个编码器和解码器，实现实时的风格化任务。此后，大量工作 [13,15] 基于这种框架设计风格迁移策略来提升风格迁移的效果。

2.2 ViTs 特征表达

目前 ViTs [3] 在计算机视觉领域的受关注度如火如荼，相比于卷积神经网络 [6,12]，ViTs 的特征表达能力明显有所优势。而不同训练方式的 ViTs，其特征表达能力也各有差异。传统的 ViTs 网络 [3] 训练通常采用有监督的方式，而目前使用半监督方式训练的 ViTs，如 CLIP [11] 等在跨模态等方面更具优势。此外，无监督训练的 ViTs(DINO [2]) 在结构特征提取上能力出色。Amir 等人 [1] 使用 DINO 的 Keys 作为特征提取，并将其应用于许多具有挑战性的无监督视觉任务中，大量惊人的结果表明了 DINO 对高级语义信息捕捉等方面的表现出色。

由于现有的风格迁移方法在结构语义信息上捕捉不足，Splice 方法着重针对这一缺点提出了视觉外观转移方法策略，利用 DINO 高级语义信息的抓取能力，以 DINO 的特征构建风格指导，实现了更高质量的风格迁移效果。

3 DINO-ViT 的结构与外观特征表达

DINO-ViT [2] 具有捕捉高级语义信息的能力，本章基于 Splice 文章对图像结构与外观特征的选择思路分析 DINO 的不同特征的表达能力。其中着重分析 DINO 的 Keys 特征对图像结构的表达以及 CLS Tokens 特征对图像外观的表达。

3.1 DINO-ViT 的结构特征表达

使用 DINO 的 Keys 特征来表达图像的结构 [1]，首先通过 PCA 特征可视化与特征逆映射的方式分析其可行性与有效性。

PCA 特征可视化是对提取的特征按通道维度进行主成分分析，将特征维度压缩至三通道，并将其规范化作为 RGB 颜色的表示。图1第 3 行展示了 PCA 特征可视化的结果，相同的像素颜色代表具有相近的语义信息。由此看出 DINO 的 Keys 特征对结构的语义信息抓取能力

较好。与 [1] 中的有监督 ViTs 的特征以及第5章图6的半监督 ViTs 特征效果相比，明显体现了 DINO 的结构语义信息抓取能力突出。此外，通过特征逆映射的方式分析其结构表达能力，图1第 2 行展示了对 DINO 第 11 层的 Keys 特征逆映射的效果。由此可见其逆映射结果与原始图像几乎一致。其中，特征逆映射方法是采样一个固定的噪声，作为可训练的生成器网络的输入，然后对生成器的输出与原始图像分别提取 DINO 的 Keys 特征，并计算二者之间的 L2 损失来优化生成器，最后以生成器输入固定噪声的输出结果作为逆映射结果。优化的目标函数为

$$\arg\min_{\theta} ||\phi(F_\theta(z)) - \phi(I)||_2 \tag{1}$$

其中，F_θ 为可训练生成器，z 和 I 分别为固定噪声和原始图像，$\phi(\cdot)$ 为 DINO 的 Keys 特征。

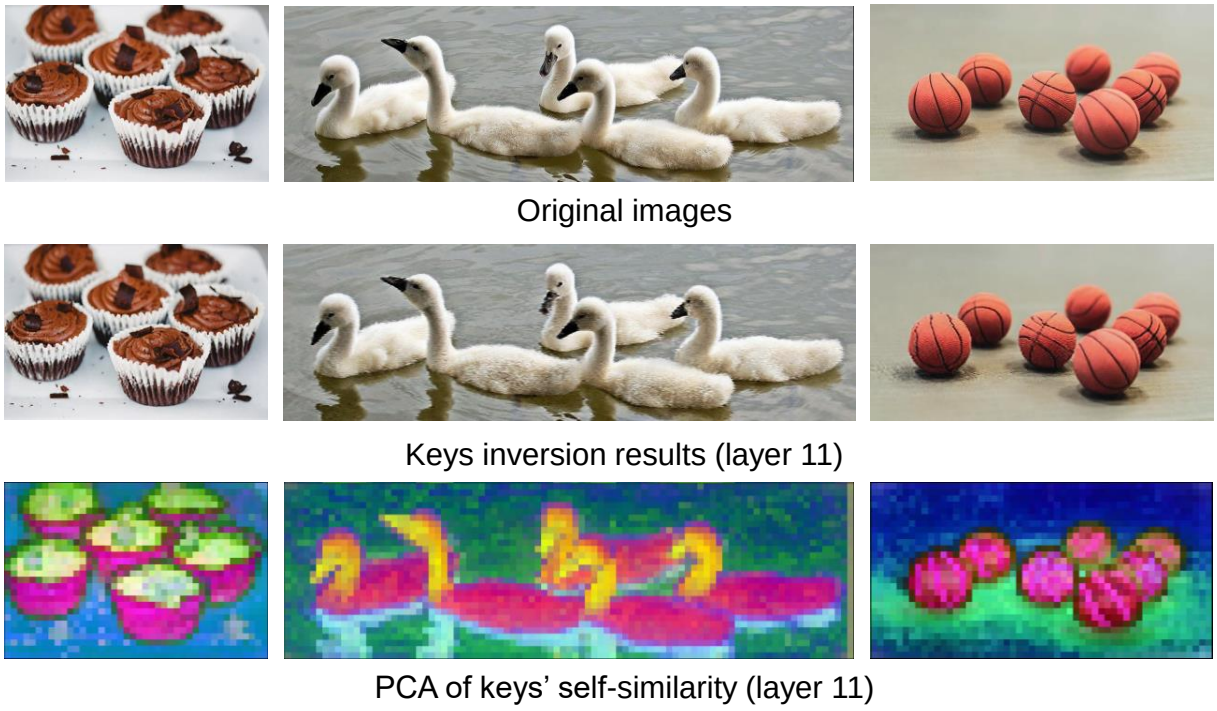

图 1. DINO Keys 特征对结构的特征表达

3.2 DINO-ViT 的外观特征表达

对于 DINO 的外观特征表达，同样采样特征逆映射的方式进行。使用公式1并将其中的 $\phi(\cdot)$ 特征替换为 CLS Tokens 特征，以同样的形式进行训练优化。图2展示了不同层的 CLS Tokens 特征的逆映射结果。图中当特征层较低时，其逆映射结果表示的更多是原始图像的局部纹理，而随着层数的增加，逆映射的结果更加能看出与原始图像相似的语义。由此可知，DINO 的 CLS Tokens 特征随着层数的增加，其所包含的全局外观语义信息更加丰富。

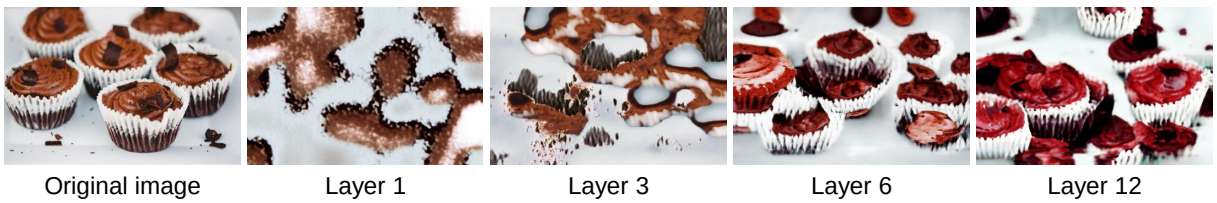

图 2. 不同 ViT 层的 CLS Tokens 特征逆映射

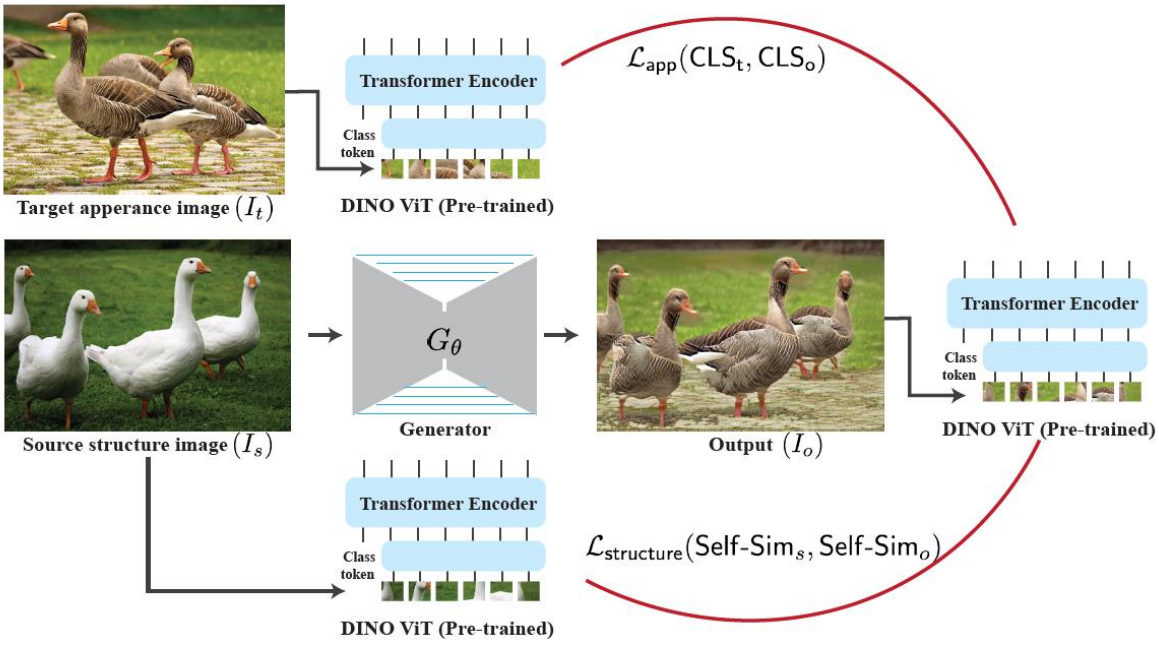

图 4. Splice 方法

另外，对最深层的 CLS Tokens 特征重复进行多次逆映射。图3展示了 3 次逆映射得到的不同结果。即便不同的运行得到的结果中结构位置有所区别，但都能看出其中所表示的物体内容与原始图像的相似。由此可知 CLS Tokens 特征抓取了全局的外观信息之外，还体现了对物体表达的多样性。

图 3. 最深层的 CLS Tokens 特征逆映射的重复执行结果

4 Splice 迁移方法

图4展示了 Splice 方法的整体框架。Splice 方法以 UNet 的网络结构构建了一个生成器 G_θ，其以目标的结构图像 I_s 作为输入，期望输出一个具有与目标结构图像一致结构以及与目标外观图像 I_t 相似的结果图像 $I_o = G_\theta(I_s)$。训练该生成器的损失函数中，Splice 方法设计了结构一致损失、外观相似损失。其中，结构一致损失为输出图像与目标结构图像的 DINO Keys 特征的自相似度计算的 L2 损失，表示为

$$\mathcal{L}_{struct} = ||S^L(I_s) - S^L(I_o)||_2 \tag{2}$$

其中，S^L 为第 L 层的 DINO Keys 特征的自相似度。与其类似，外观相似损失为输出和图像与目标外观图像的 CLS Tokens 特征计算的 L2 损失，表示为

$$\mathcal{L}_{app} = ||t^L_{[cls]}(I_t) - t^L_{[cls]}(I_o)||_2 \tag{3}$$

其中，$t^L_{[cls]}$ 为第 L 层的 CLS Tokens 特征。另外，为了进一步提升生成结果的质量，额外设计了一个 ID 损失，以输出图像与目标外观图像的 DINO Keys 特征计算的 L2 损失来表示：

$$\mathcal{L}_{id} = ||K^L(I_t) - K^L(G_\theta(I_t))||_2 \tag{4}$$

其中 K^L 为第 L 层的 DINO Keys 特征。在本文中 L 均为 11，即最深层。

因此，Splice 方法的总损失为

$$\mathcal{L}_{total} = \mathcal{L}_{struct} + \alpha\mathcal{L}_{app} + \beta\mathcal{L}_{id} \tag{5}$$

由于模型训练中，训练数据仅为输入的目标结构图象与目标外观图像，因此对训练图像进行数据增强是必要的。训练中，采用随机裁剪、颜色扰动、随机翻转、随机模糊化四个增强操作的随机组合对训练图像进行增强，其中目标外观图像只采取随机翻转和随机裁剪操作。图5展示了部分随机数据增强的效果。

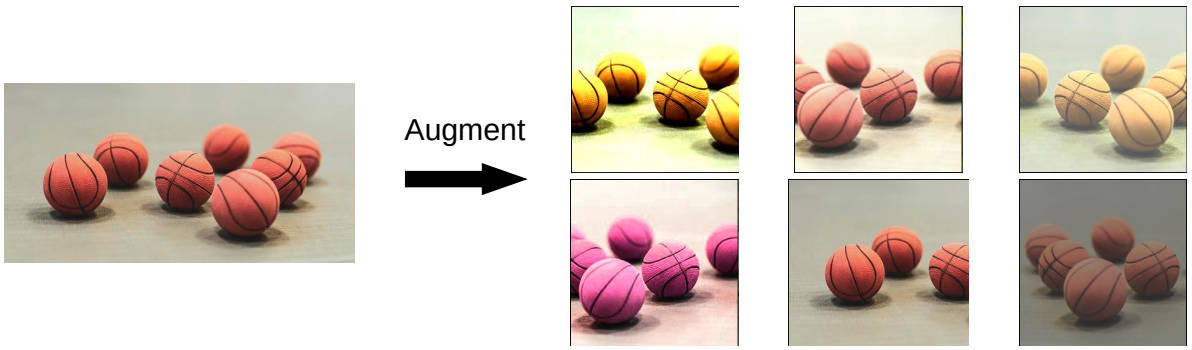

图 5. 训练数据增强

5 实验结果与分析讨论

5.1 复现细节

5.1.1 与已有开源代码对比

Splice 方法 [14] 的代码已开源[1]，本文复现代码详见 Github 仓库[2]。复现代码中，为了与预训练模型的网络参数适配，ViTs 的网络结构代码从 timm[3]中拷贝；另外，ViTs 特征提取的实现中部分参考了 Splice 开源代码。相比开源代码，在本文复现工作中新增了对 CLIP-ViT 的特征提取以展开特征可视化的对比分析；另外重写并精细化了 UNet 网络结构，使得生成结果的质量效果稍微优于开源代码。

5.1.2 实验环境搭建

本文方法复现在 Windows 10 系统下进行，使用 PyTorch 1.13.0 与 Torchvision 0.14.0 框架实现，使用 CUDA 版本为 11.6。

[1] https://github.com/omerbt/Splice

[2] https://github.com/Zichong-Chan/Splice-transfer

[3] https://github.com/rwightman/pytorch-image-models/blob/master/timm/models/vision_transformer.py

5.2 复现结果

5.2.1 结构与外观特征表达

图6展示了 DINO 与 CLIP 的 Keys 特征 PCA 可视化结果，进一步体现了无监督训练的 DINO 在结构语义信息提取方面优于半监督/文本监督的 CLIP-ViT 模型。

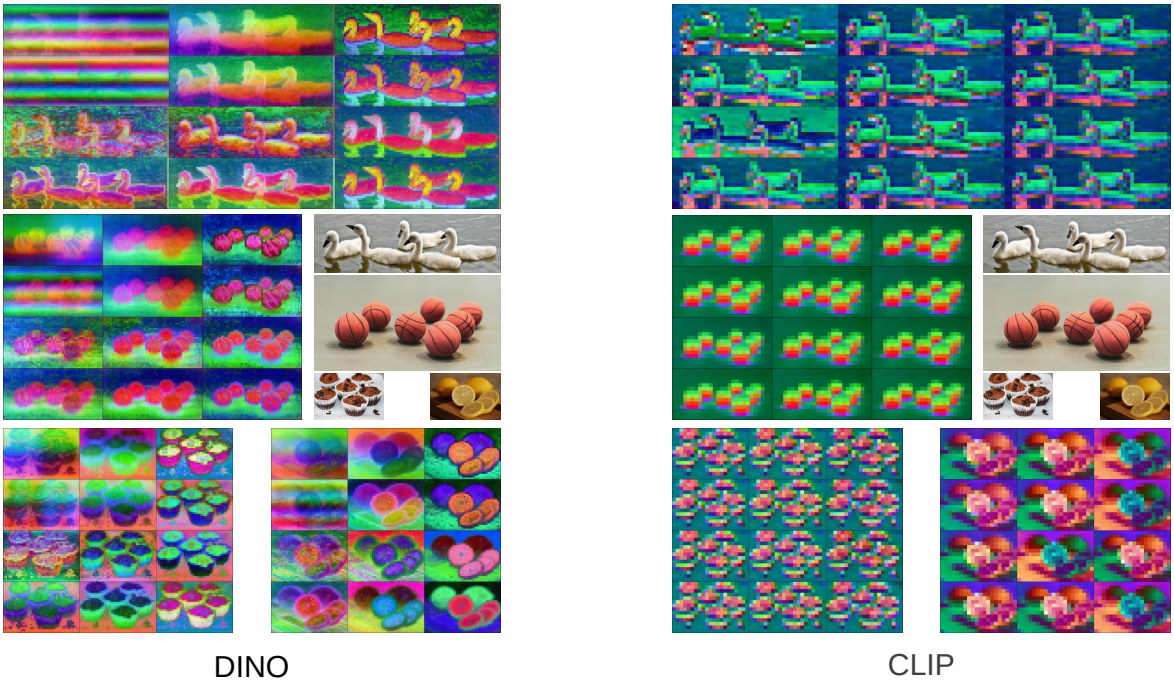

图 6. PCA 特征可视化

图7展示了更多使用 CLS Tokens 特征逆映射结果。在不同的图像样例中，CLS Tokens 都有较好的全局外观语义信息抓取，充分表明了 Splice 方法选取 CLS Tokens 特征作为图像外观表达的可行性与有效性。

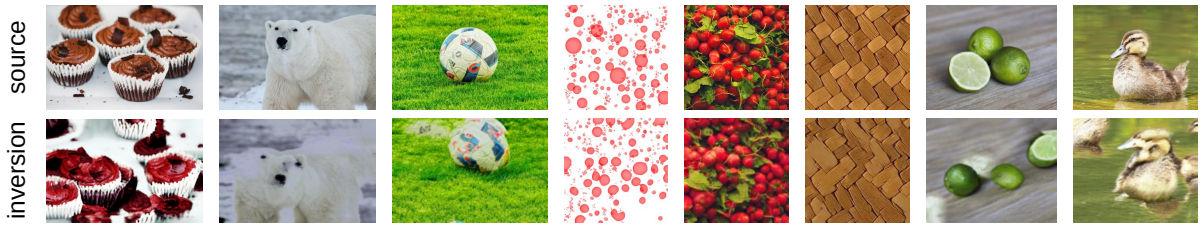

图 7. CLS Tokens 特征逆映射

图8展示了更多使用 DINO Keys 特征逆映射结果。在不同的图像样例中，DINO Keys 特征都较好地还原了与原始图像的结构与外观，在结构语义信息的抓取上，充分表明了 Splice 方法选取 Keys 特征构建图像结构表达的可行性与有效性。

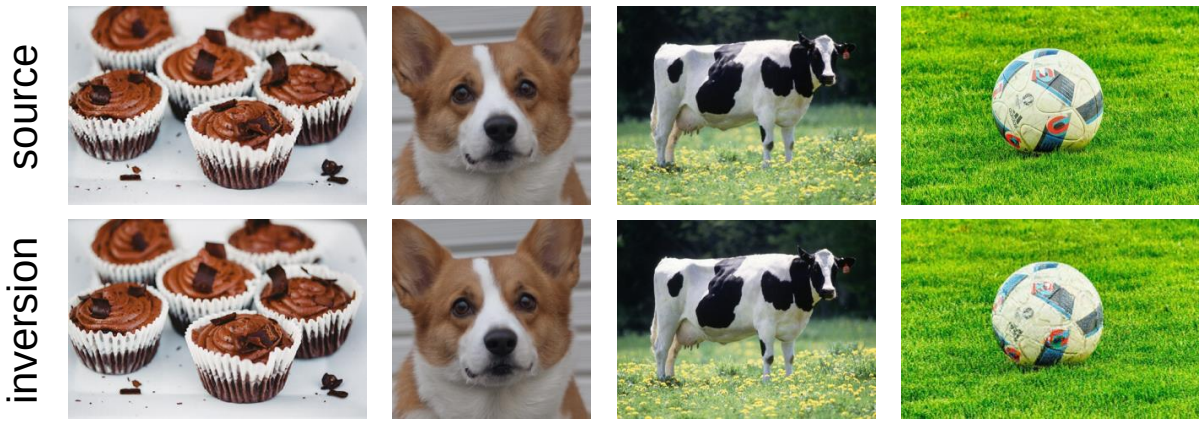

图 8. Keys 特征逆映射

5.2.2 Splice 迁移结果

图9展示了多组视觉外观转移的生成结果，其中每个三元组图像中，左边和右边的图像分别为目标结构图像与目标外观图像，中间的图像为生成结果。图中结果表明，在不同类别的语义图像中，Splice 方法都能较好地实现视觉外观的转移。

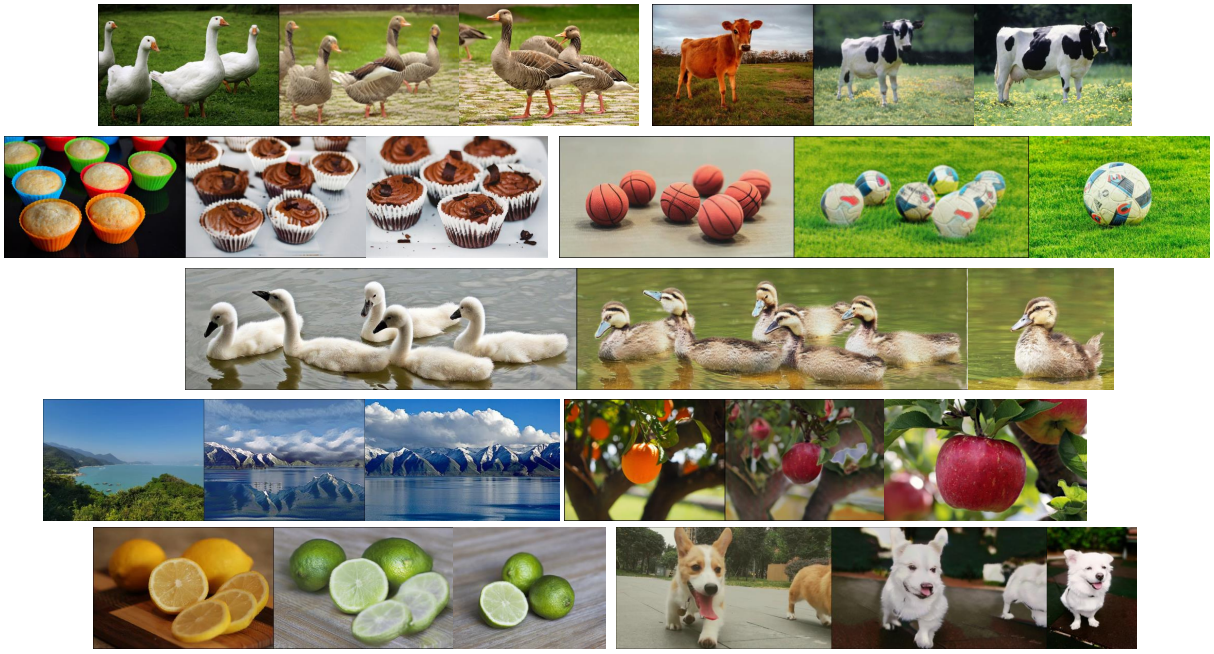

图 9. Splice 迁移结果

5.3 讨论

即便 Splice 方法能够实现较好的视觉外观转移效果，该方法仍然存在一些不足之处。

首先是训练后的生成器泛化能力不够。如图10所示，使用篮球图像 (第一行左 1) 作为目标结构图像训练生成器后，将该图像输入至生成器可以得到较好的迁移效果，然而即使同样是球类的图像，生成结果都不能得到比较理想的效果。其主要原因是训练数据中只有一个目

标结构图像，生成器对其他图像的迁移固然泛化能力有所不足。然而对于每对目标图像进行视觉外观迁移时都要训练这样的一个生成器，其成本代价过大，导致其实用性不足。

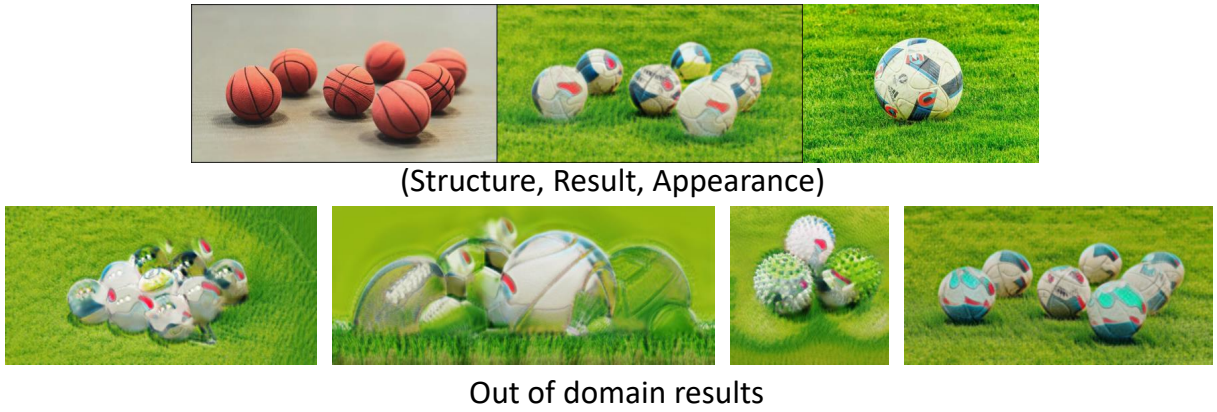

图 10. 较差的泛化能力例子

另外，图11展示了使用不同语义类别的目标图像作为训练数据时存在语义匹配失败的例子。使用域间差异较大的两个图像作为目标图像进行训练，其迁移结果不理想。其主要原因应该是 DINO 自监督训练的预训练模型是在 ImageNet 数据集上进行，其训练策略本身会使得模型具备一定的分类能力，而当图像的语义类别不一致时，语义匹配失败也难以避免。

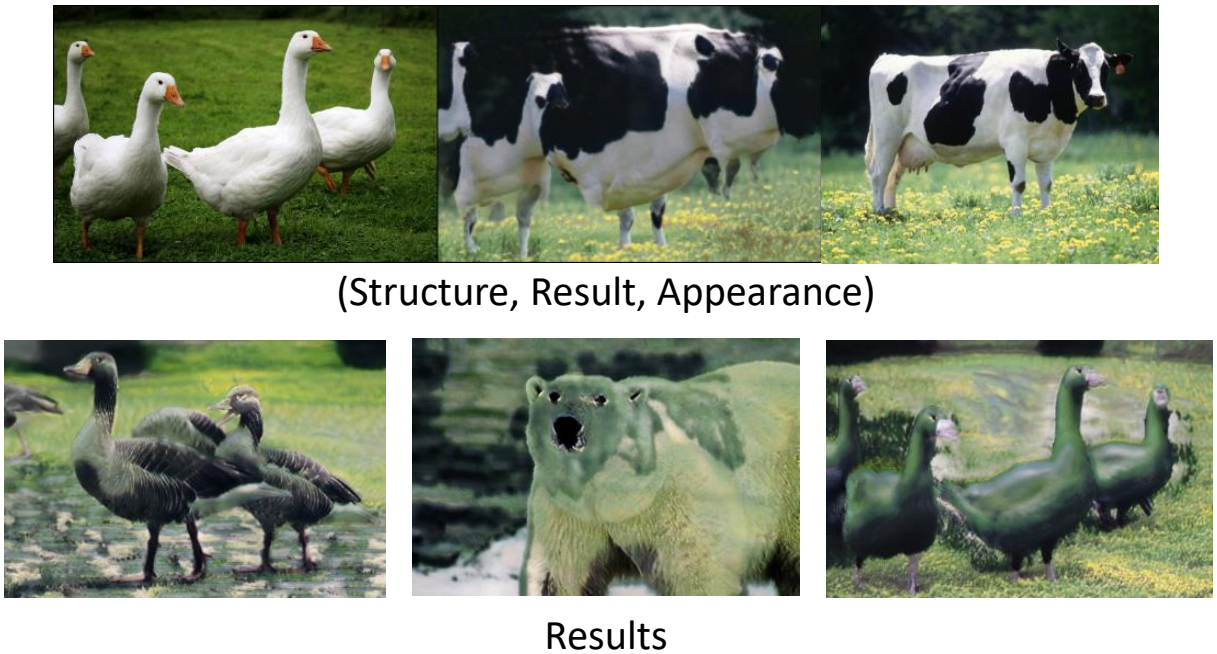

图 11. 较差的语义匹配例子

然而，即使 Splice 方法存在上述两点明显的不足，但是该方法使用 DINO 的特征构建了一个足够简单易用的迁移方法，对该领域带来了一定的价值。鉴于目前基于 UNet 与交叉注意力机制的扩散模型的成功应用，本文提出对 Splice 方法的改进思路以提高该方法的泛化能力：将基于 UNet 结构的生成器进行修改，增加交叉注意力模块，将目标外观图像的 CLS Tokens 通过编码的方式注入到生成路径中，然后使用大量数据训练一个具有任意外观与结构的生成器，将原始基于类优化的 Splice 方法改进为基于编码器的方法。

6 总结与展望

本文对 Splice 方法进行了解读，并通过特征可视化的方式认识与分析了 DINO 特征的高级语义信息表达能力，重现了 Splice 选择 DINO 的 Keys 特征和 CLS Tokens 特征作为图像结构与外观表达的可行性与有效性。另外通过方法复现，实现了基于 DINO 特征的视觉外观转移任务，并通过大量实验结果表明了复现结果的正确性，同时充分证明了 Splice 方法在一定程度上的优势。Splice 方法简单易用，在未来的工作中可以尝试第5.3节中的改进思路进一步提升该方法的泛化能力。

参考文献

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. Int. Conf. on Computer Vision*, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. on Learning Representations*, 2021.

[4] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 341–346, 2001.

[5] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 2414–2423, Las Vegas, NV, USA, June 2016. IEEE.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 770–778, 2016.

[7] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 327–340, 2001.

[8] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. Int. Conf. on Computer Vision*, 2017.

[9] Sunnie S. Y. Kim, Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Deformable style transfer. In *Proc. Euro. Conf. on Computer Vision*, 2020.

[10] N. Kolkin, J. Salavon, and G. Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 10043–10052, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. on Learning Representations*, 2015.

[13] Chunjin Song, Zhijie Wu, Yang Zhou, Minglun Gong, and Hui Huang. Etnet: Error transition network for arbitrary style transfer. In *Proc. Conf. on Neural Information Processing Systems*, pages 668–677, 2019.

[14] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proc. IEEE/CVF Conf. on Computer Vision & Pattern Recognition*, pages 10748–10757, 2022.

[15] Zhijie Wu, Chunjin Song, Yang Zhou, Minglun Gong, and Hui Huang. Efanet: Exchangeable feature alignment network for arbitrary style transfer. *Proc. AAAI Conf. on Artificial Intelligence*, pages 12305–12312, 4 2020.