专题综述

基于 Transformer 的行人重识别研究与展望

大连理工大学人工智能学院 张平平

行人重识别（Person Re-identification, Re-ID）技术的目的是依据不同时间和地点拍摄的图像或视频内容，检索场景中的特定行人。由于该技术在安全社区、智能监控和刑事侦查等前沿应用中的重要性，近年来已经成为计算机视觉领域研究的热点问题之一。然而，在现实复杂场景下，由于其易受到摄像机视角变化、行人姿态变化、物体区域遮挡、图像低分辨率、行人图像未对齐等诸多因素的影响，精确高效的行人重识别仍是一项极具挑战性的研究课题。

在过去的十余年里，行人重识别的研究取得了很大的进展，并产生了一系列的相关任务，如基于图像的行人重识别(Image-based Re-ID)、基于视频的行人重识别 (Video-based Re-ID)、行人搜索 (Person Search) 等。这些任务的完成离不开图像或者视频信息的鲁棒特征表示和检索度量(策略)的提升。如何获得更加鲁棒的视觉表征是制约着行人重识别性能提升的关键。早期的行人重识别算法主要关注手工特征的提取和相似性度量的设计。随着深度学习技术的发展，越来越多的工作聚焦端到端地学习更具判别性的深度特征，主要瞄准设计更加复杂的深度卷积神经网络（Convolutional Neural Network, CNN)。然而，深度 CNN 是通过逐层堆叠卷积操作实现的，而卷积是一种局部操作，一个卷积层通常只会建模邻域像素之间的关系，并不能实现信息的全局建模，这严重制约着行人重识别的性能。近期，基于 Transformer 的模型[1]无论是在自然语言处理领域还是在计算机视觉领域均取得了优异的表现，其核心原因在于 Transformer 是基于自注意力的全局操作，可以建模所有元素之间的关系，从而普遍提升模型的全局感知能力。得益于此，已经有一些工作尝试使用 Transformer 模型完成行人重识别，并取得了优异的性能。本文将重点介绍近期基于 Transformer 的行人重识别相关研究进展和未来发展趋势。

一、基于纯Transformer的行人重识别

纯 Transformer 模型在行人重识别领域的代表性工作是发表在 ICCV2021 上的 TransRe-ID[2]，该工作直接借鉴 ViT[3]模型处理图像数据的思路，将行人图像分成多个图像块(例如 16x16 像素大小)，并把这些图像块作为序列输入标准的 Transformer 编码器中，提取行人的判别性特征用于行人重识别。此外，为了进一步增强特征的鲁棒性，该工作还设计了两个新的模块：(i)拼图模块(Jigsaw Patch Module, JPM)，通过移动和混洗操作对图像块的嵌入进行重新排列，使区域覆盖范围更加多样化，提高了行人重识别能力。(ii)侧信息嵌入(Side Information Embeddings, SIE)，通过插入可学习嵌入来融合非视觉线索，从而减轻对相机/视图变化的特征偏差。该工作的具体框架如图 1 所示。

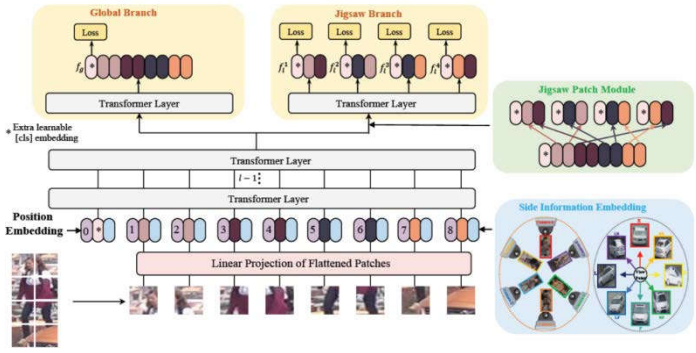

图 1　TransRe-ID 模型框架

总体而言，TransRe-ID 框架对 Transformer 的应用比较简单直接，没有考虑行人本身的结构化特点和空间的连续性，因而在行人出现遮挡或者不对齐时表现不佳。为此，学术界开展了一些列的改进工作。如，Sharma 等人[4]提出了一种局部感知 Transformer(Locally Aware-Transformer, LAT)，将全局增强的局部特征聚合到集成分类器中实现行人重识别，其具体框架如图 2 所示。为了提升对遮挡行人的表示能力，Zhao 等人[5]设计了一种基于局部特征的 Transformer(Partial Feature Transformer, PFT)用于行人重识别。其主要贡献是构建了图像块全维增强模块、融合重建模块和空间切片模块，显著提高了遮挡下的行人重识别性能。为了处理行人图像不对齐问题，Zhu 等人[6]首次在 Transformer 体系结构中引入了一种对齐方案，并提出了自动对齐的 Transformer(Auto-Aligned Transformer, AAformer)用于在图像块级别上自动定位行人和非行人部件。同时，将部件对齐集成到自注意模块中，输出的部件特征可以直接用于行人检索，其具体框架如图 3 所示。大量的数值实验验证了所学部件的有效性。此外，为了克服行人重识别任务对大量标注数据的依赖，Cao 等人[7]结合多标签分类法，将 ViT 应用于无监督行人重识别任务。实验结果表明，增强的 ViT 模型的性能普遍优于传统方法和大多数基于 CNN 的方法。为了降低初始伪标签的噪声，Xia 等人[8]基于 Transformer 设计了特征提取模型 Trans-Encoder。与传统的 CNN 相比，Trans-Encoder 提取的特征对跨域迁移具有更强的鲁棒性。在此基础上，可以提高特征聚类的置信度。

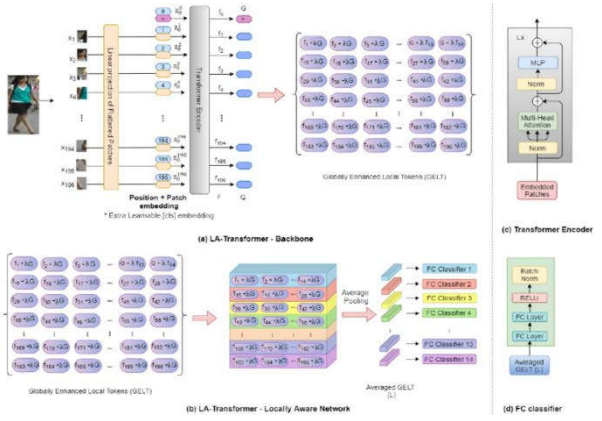

图 2　基于局部感知 Transformer 的行人重识别框架

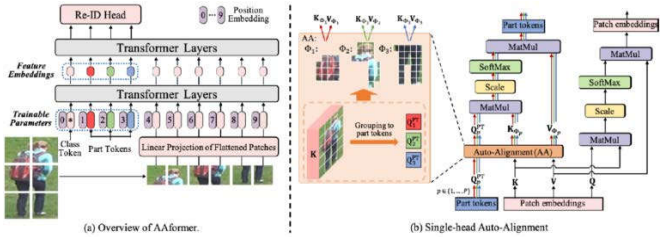

图 3　基于自动对齐 Transformer 的行人重识别框架

尽管基于纯 Transformer 的行人重识别方法取得了较大的成功，但是这类方法仍存在一些明显问题，如模型复杂度高、难以快速训练、在一定程度上忽略了行人图像的空间结构信息、无法在主流计算机设备上部署等。因而，基于纯 Transformer 的行人重识别模型仍有较大的提升空间。

二、CNN和Transformer耦合的行人重识别

在行人重识别任务中，利用判决性部件特征往往能带来额外的准确率提升，因而如何更好地学习行人的部件特征也是行人重识别成功的关键。众所周知，CNN 模型更注重局部特征的提取，因而天然地适合提取部件信息。然而，受制于有限的接受域，通常 CNN 提取的特征并不具有较强的全局特性，这又制约着 CNN 在行人重识别这一语义检索类任务上效能的发挥。而 Transformer 模型对空间和序列数据具有很强的长距依赖关系建模能力，天然地适合提取全局语义信息。因而，如何充分结合这两类模型的优势，构建更加精确和鲁棒的行人重识别模型也是当前研究的热门方向之一。而目前基于 CNN 和 Transformer 耦合的行人重识别方法可以大致分为如下三类：

2.1. 底层 CNN+高层 Transformer 架构

针对 CNN 和 Transformer 提取特征的不同，最直接的融合提升方法就是首先使用 CNN 提取行人的底层视觉特征，然后利用 Transformer 的全局建模能力汇聚得到具有高级语义信息的检索特征。沿着这一思路，目前大部分基于 Transformer 的行人重识别算法均取得了优于主流 CNN 模型的性能。如在基于图像的行人重识别中，Zhou 等人[9]首先利用 ResNet-50 提取多尺度视觉特征，然后通过 Transformer 编码器聚合查询样本

的 k 近邻上下文信息，最终设计了一种重排序网络来预测查询样本和排名靠前的邻居样本之间的相关性。在 6 个常用的行人和车辆重识别数据集上进行了实验，验证了该方法的有效性。此外，由于目标行人经常被各种障碍物或其他人遮挡，为了解决这些问题，Li 等人[10]提出了一种端到端的部件感知 Transformer(Part-Aware Transformer)，通过 CNN 提取图像视觉信息，同时构建像素上下文 Transformer 和部件原型 Transformer 挖掘不同的部件信息，实现了对被遮挡行人的不同部位的重识别，其具体框架如图 4 所示。Jia 等人[11]利用 CNN 和 Transformer 架构，通过对被遮挡行人图像的局部特征进行全局推理，实现了无对齐的行人重识别。为了实现跨域的行人重识别，Waseem 等人[12]提出了一种使用 CNN 混合视觉 Transformer 的域适应方法，并将聚类损失函数和广泛使用的三重损失函数合并在一起，改善了现有的无监督领域自适应行人重识别方法的性能。

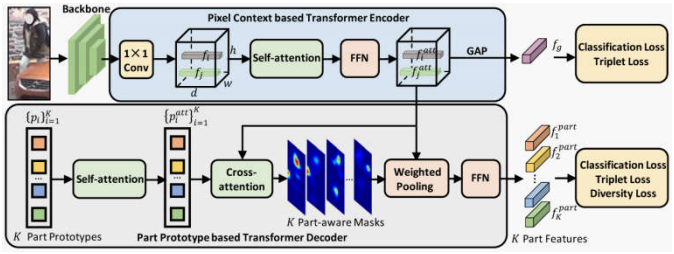

图 4 基于部件感知 Transformer 的行人重识别框架

针对基于视频的行人重识别，为了获取更丰富的感知信息和提取更全面的视频表示，Liu 等人[13]提出了一种基于多视角学习的三叉 Transformer(Trigeminal Transformer，TMT)框架，如图 5 所示。具体来说，该工作首先将原始视频数据通过 CNN 联合转换为空间、时间和时空域特征，然后利用三种自我视图的 Transformer 来增强空间、时间和时空域的信息。此外，还提出了一个交叉视图 Transformer 来聚合多视图特征，以实现全面的视频表示。实验结果表明，在公开的三个数据集上，该方法可以获得比其他同时期最先进的方法更好的性能。此外，He 等人[14]发现有效地提取多尺度细粒度特征并构建它们之间的结构交互是基于视频行人重识别成功的关键。因此，他们提出了一个混合框架，即密集交互学习(Dense Interaction Learning, DenseIL)，它综合利用了 CNN 和 Transformer 架构的主要优点。如图 6 所示，DenseIL 包含一个 CNN 编码器和一个密集交互 Transformer 解码器。CNN 编码器负责有效地提取空间特征，而 Transformer 解码器被设计成密集地模拟跨帧的时空固有交互作用。

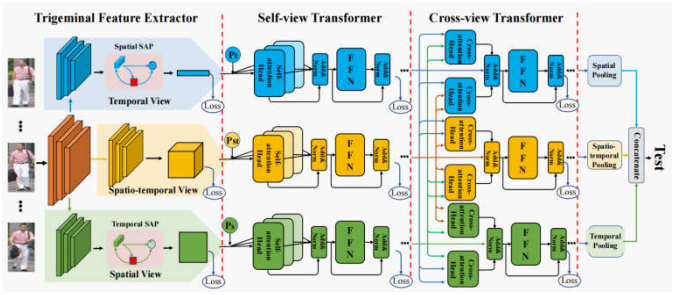

图 5 基于三叉 Transformer 的视频行人重识别框架

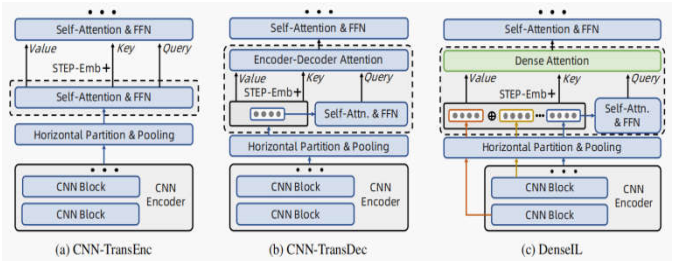

图 6 基于密集交互学习的视频行人重识别框架

2.2. 深度多层次耦合

前面所述的行人重识别方法主要是将 CNN 和 Transformer 作为独立的功能模块实现鲁棒的特征表示。事实上，它们忽略了 CNN 和 Transformer 均是层次化的表示模型，不同的卷积层和 Transformer 层可以表示不同的抽象信息。为了实现更加丰富和充分的信息融合，一些研究者尝试从深度多层次耦合的角度实现 CNN 和 Transformer 的互补，并提升行人重识别的性能。如，Tahir 等人[15]直接采用了三种不同的 CNN 网络结构，将 Transformer 模块插入到 CNN 的不同层，构建了新的主干网实现行人重识别。Zhang 等人[16]利用 CNN 和 Transformer 的层次化特点，提出了一种基于层次化聚合的 Transformer(Hierarchical Aggregation Transformer, HAT)框架用于图像的行人重识别，其结构如图 7 所示。为了解决行人裁剪后的轨迹时空偏差，Liu 等人[17]利用改进的 Transformer 构建了逐层的由粗到细轴向注意网络。该工作不仅能显著降低计算量，而且无需考虑空间和时间对齐以及数据集噪声的影响。

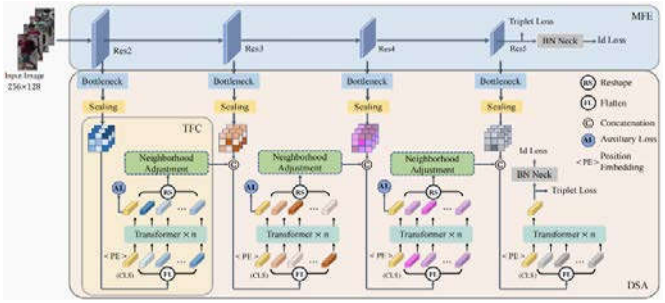

图 7 基于层次化聚合 Transformer 的行人重识别框

高性能的行人重识别要求模型同时关注行人的全局轮廓和局部细节。为了提取更具有代表性的特征，一种有效的方法是利用具有多个分支的深度模型。然而，大多数基于多分支的方法都是通过部分骨干结构的重复来实现的，通常会导致计算量的显著增加。为此，Zhang 等人[18]借鉴目前流行的特征金字塔网络(Feature Pyramid Network, FPN)，同时使用 Transformer 结构从不同的网络层提取全局特征，并将它们聚合成一个双向金字塔结构应用于行人重识别任务中。实验表明该方法取得了较好的识别效果。

2.3. 额外信息引导耦合

在行人重识别的现实应用中，额外的关联信息如摄像机信息、行人姿态、语言描述、属性标签等也能起到关键的作用。因而，如何利用这些额外信息或者在这些信息的引导下实现 CNN 和 Transformer 的耦合，进而提升模型的性能也是目前研究的一大趋势。为了从跨摄像机非配对训练数据中学习摄像机的不变表示，Ge 等人[19]提出了一种基于摄像机引导的 Transformer 行人重识别框架。该工作通过 CNN 变换伪跨摄像机正特征对，最小化伪特征对的距离，从摄像机特定的特征分布中挖掘出跨摄像机的自监督信息。此外，同步利用 Transformer 实现局部特征的自动定位和提取，进而实现超远距离的行人重识别。鉴于行人的姿态信息在识别过程中也扮演着重要作用，Ma 等人[20]提出了一种姿态引导的部件间和部件内关系 Transformer(Pose-guided Inter- and Intra-part Relational Transformer, PIRT)，其框架如图 8 所示。该工作通过引入 CNN 来生成姿态掩码，Transformer 来建立局部感知的长距相关性，实现了模型耦合性能的提升。类似地，Wang 等人[21]构建了基于姿态引导的特征解纠缠方法。该工作利用 Transformer 提取全局视觉特征，在姿态引导的特征聚合模块中利用匹配和分布机制，初步将姿态信息与图像块信息分离，实现了遮挡条件下的行人重识别。

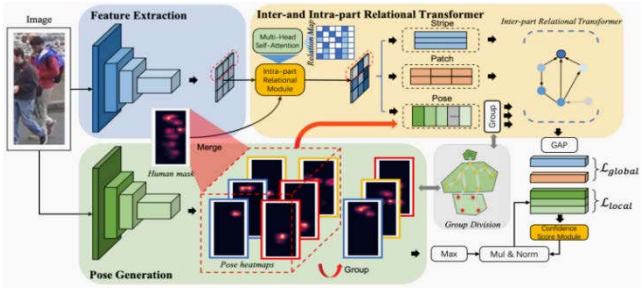

图 8 基于姿态引导耦合的行人重识别框架

预训练是计算机视觉的主导范式。为了寻找传统的预训练替代方法，Xiang 等人[22]提出了一种基于文本引导的行人重识别骨干网络预训练方法。该方法使用 CNN 提取图像的视觉特征，同时利用 Transformer 从文本标注中学习视觉表示，从而实现了 CNN 和 Transformer 的耦合学习。在基准测试集上进行的实验表明，与在 ImageNet 上预训练的模型相比，该方法可以取得极具竞争力的性能，揭示了它在行人重识别任务上的潜力。

三、总结与展望

本文介绍了近期基于 Transformer 的行人重识别方法，包括基于纯 Transformer 的模型以及基于 CNN 和 Transformer 耦合的模型。相关工作表明，有效地构建 CNN 和 Transformer 耦合方法对于提升行人重识别的性能至关重要。未来的发展方向包括以下几个方面：如何将 Transformer 这一全局/长距建模能力很强的模型应用于其他视觉模态以及与服装无关的生物特征上，从而实现多模态、跨媒体、超长时的行人重识别；如何将行人重识别与目标检测、多目标跟踪、行人分割等相关任务进行联合建模，从而发挥多任务学习的优势，促进视觉任务的有机融合;如何设计精确且高效的轻量化 Transformer 模型以及寻找经济和高效的训练方式(包括无监督、半监督、自监督等)；如何开发面向任务和用户的可解释 Transformer 网络，实现模型的可解释性。

责任编委 王金甲

参考文献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. NeurIPS 2017.
[2] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification[C]. ICCV 2021.
[3] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]. ICLR 2021.
[4] Sharma C, Kapil S R, Chapman D. Person re-identification with a locally aware transformer[J]. arXiv:2106.03720, 2021.
[5] Zhao Y, Zhu S, Wang D, et al. Short Range Correlation Transformer for Occluded Person Re-Identification[J]. arXiv:2201.01090, 2022.
[6] Zhu K, Guo H, Zhang S, et al. Aaformer: Auto-aligned transformer for person re-identification[J]. arXiv:2104.00921, 2021.
[7] Cao G, Jo K H. Unsupervised Person Re-Identification with Transformer-based Network for Intelligent Surveillance Systems[C]. ISIE 2021.
[8] Xia L, Yu Z, Ma W, et al. Refining Pseudo Labels for Unsupervised Domain Adaptive Person Re-Identification[J]. IEEE Access 2021.
[9] Zhou Y, Wang Y, Chau L P. Moving Towards Centers: Re-ranking with Attention and Memory for Re-identification[J]. arXiv:2105.01447, 2021.
[10] Li Y, He J, Zhang T, et al. Diverse part discovery: Occluded person re-identification with part-aware transformer[C]. CVPR 2021.
[11] Jia M, Cheng X, Lu S, et al. Learning Disentangled Representation Implicitly via Transformer for Occluded Person Re-Identification[J]. IEEE TMM 2022.
[12] Waseem M D, Tahir M A, Durrani M N. Hybrid Vision Transformer for Domain Adaptable Person Re-identification[C]. ICCCI 2021.
[13] Liu X, Zhang P, Yu C, et al. A Video Is Worth Three Views: Trigeminal Transformers for Video-based Person Re-identification[J]. arXiv:2104.01745, 2021.
[14] He T, Jin X, Shen X, et al. Dense Interaction Learning for Video-based Person Re-identification[C]. CVPR 2021.
[15] Tahir M, Anwar S. Transformers in Pedestrian Image Retrieval and Person Re-Identification in a Multi-Camera Surveillance System[J]. Applied Sciences 2021.
[16] Zhang G, Zhang P, Qi J, et al. Hat: Hierarchical aggregation transformers for person re-identification[C]. ACM MM 2021.
[17] Liu C T, Chen J C, Chen C S, et al. Video-based Person Re-identification without Bells and Whistles[C]. CVPR 2021.
[18] Zhang S, Yin Z, Wu X, et al. FPB: Feature Pyramid Branch for Person Re-Identification[J]. arXiv:2108.01901, 2021.
[19] Ge W, Pan C, Wu A, et al. Cross-Camera Feature Prediction for Intra-Camera Supervised Person Re-identification across Distant Scenes[C]. ACM MM 2021.
[20] Ma Z, Zhao Y, Li J. Pose-guided Inter-and Intra-part Relational Transformer for Occluded Person Re-Identification[C]. ACM MM 2021.
[21] Wang T, Liu H, Song P, et al. Pose-guided Feature Disentangling for Occluded Person Re-identification Based on Transformer[J]. arXiv:2112.02466, 2021.
[22] Xiang S, Zhang Z, Guan M, et al. VTBR: Semantic-based Pretraining for Person Re-Identification[J]. arXiv:2110.05074, 2021.

张平平

大连理工大学人工智能学院副教授。主要研究方向是：计算机视觉、深度学习。目前已经在计算机视觉和人工智能相关领域的国内外著名学术期刊和会议发表论文 40 余篇，担任多个国际学术期刊和会议的审稿人及程序委员会委员。

Email: zhpp@dlut.edu.cn

热点追踪

抗姿态和遮挡的人脸生成与识别

重庆大学 段青言　张 磊

姿态变化和面部遮挡是影响人脸识别最主要的两个因素。对于姿态变化，通常采用抗姿态的特征表达和基于生成对抗网络（Generative Adversarial Net，GAN）的人脸正面化这两种方式进行解决。对于面部遮挡，基于 GAN 模型的人脸修复方法也层出不穷，这些方法更多地关注正面或近正面的人脸面部结构以及像素细节，而非身份判别性。可见，姿态变化和面部遮挡通常被当作两个单独的任务来分别加以解决。然而，在实际生活中，这两种情况常常同时发生，且逐渐演变成为一种富有挑战且有待研究的问题。如图 1 中，第一行和第二行的输入图片姿态角分别为 45°和 60°，其中，(a) 为输入的侧面遮挡人脸，(b) 为我们提出的 BoostGAN 的生成结果，(c) 和 (d) 分别为现有两种人脸正面化方法的生成结果，(e) 为真实的正面人脸。从图 1 可以看出，当侧面人脸出现部分遮挡时，(c) 和 (d) 作为仅针对姿态变化的人脸正面化方法出现了不同程度的生成误差。

解决姿态变化和面部遮挡的混合问题，直接的想法是分两步进行处理，即先采用人脸修复方法去遮挡，然后人脸正面化。然而，两步法容易产生较大的误差且依

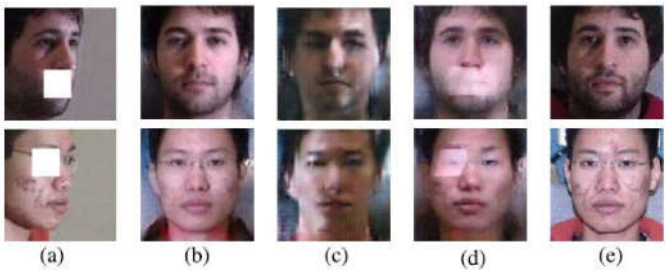

图 1 遮挡对现有人脸正面化方法的影响

赖大量训练样本。我们提出的 BoostGAN 是一种端到端的集成式生成模型，如图 2 所示，采用多张局部遮挡的图片作为模型的输入，来完备身份和纹理信息。该网络由一个深度的编-解码器（即粗糙网络）和一个浅层的集成网络（即精细网络）组成。粗糙网络用于在多重遮挡和大姿态变化的人脸图像上实现粗糙的正面化和去遮挡生成。而精细网络旨在通过集成多个中间输出的互补信息，生成干净、正面的人脸图像并保持身份特异性。

更进一步，考虑到噪声作为先验知识被广泛应用至图像修复中，以及人脸修复和人脸正面化两个任务之间的协同作用，我们提出一种遮挡掩模引导下的两阶段生成对抗网络（TSGAN），如图 3 所示。该网络主要包含

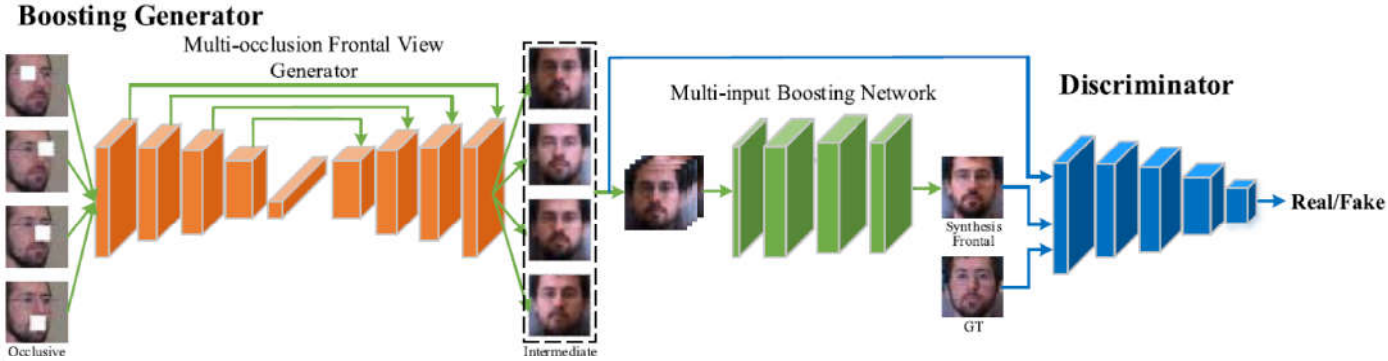

图 2 端到端由粗糙到精细生成的 BoostGAN 框架

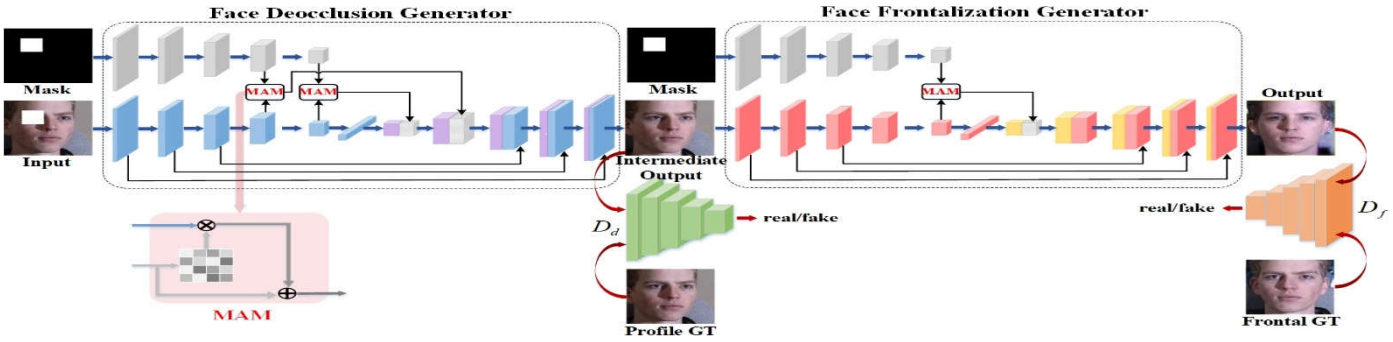

图 3 两阶段生成对抗网络（Two-Stage GAN，TSGAN）

3 个模块，即人脸去遮挡模块、人脸正面化模块和掩模注意力模块（Mask Attention Module, MAM）。前两个模块分别被设计用于不同的阶段，而 MAM 则在两个阶段中均被部署。在第一个阶段，作为一种重要的先验知识，引入遮挡掩模来拟合输入图像中的噪声，作为辅助信号帮助 TSGAN 完成人脸修复。MAM 使得人脸去遮挡模块更多的关注和更好地填充侧面人脸图像上的“空洞”，如图 4 所示。在第二阶段，由第一阶段生成的无遮挡侧面人脸作为人脸正面化模块的输入，通过 MAM 进一步获得最终逼真的正面图像。值得注意的是，TSGAN 模型是一个端到端的结构。此外，为了更有效地分别监督两个阶段保持身份的一致性和提高身份相关特征的判别性，提出针对去遮挡和正面化的双重三元损失来联合训练 TSGAN 的两个阶段。

在约束和非约束的人脸图像数据集上，定量和定性

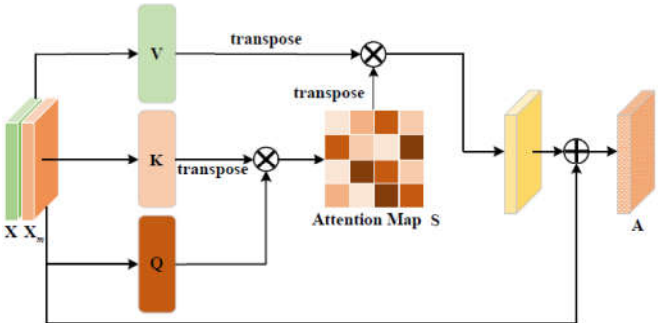

图 4 遮掩模注意力模块（MAM）

实验表明了提出的两个模型 BoostGAN 和 TSGAN 对遮挡、侧面人脸生成和识别的优越性，达到 SoTA。

以上 2 个成果分别被国际期刊 IEEE Transactions on Neural Networks and Learning Systems (2020) 和 IEEE Transactions on Circuits and Systems for Video Technology (2021)接收。

责任编委 储珺

段青言

2021 年 6 月博士毕业于重庆大学，现为重庆邮电大学讲师，主要研究方向为人脸识别、人脸生成、深度学习。
Email: duanqy@cqupt.edu.cn

张磊

重庆大学教授，博士生导师，IEEE/CCF 高级会员。主要研究方向为开放环境视觉感知、深度学习、迁移学习等。
Email: leizhang@cqu.edu.cn