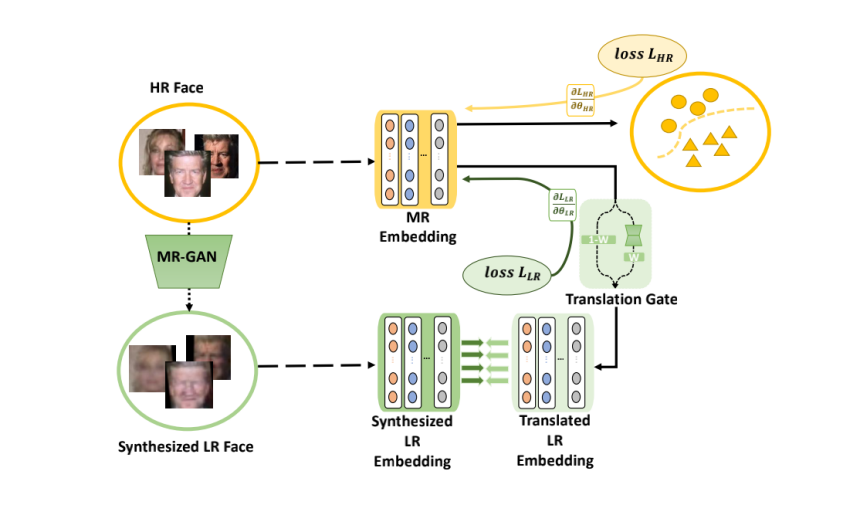
**Generate to Adapt: Resolution Adaption Network for Surveillance Face Recognition**

**Abstract**

尽管深度学习技术已大大改善了人脸识别，但由于训练数据有限和域分布的差距，**无约束的监视环境**的人脸识别仍然是一个尚未解决的挑战。先前的方法大多匹配不同域中的低分辨率和高分辨率人脸，这在普通识别场景中往往会破坏原始特征空间。为避免此问题，我们提出了分辨率自适应网络（RAN），其中包含**多分辨率生成对抗网络（MR-GAN）**，然后是特征自适应网络。**MR-GAN学习多分辨率表示，并随机选择一种分辨率以生成可避免下采样人脸产生伪影的逼真的低分辨率（LR）人脸**。开发了一种新颖的带有翻译门的特征自适应网络，**将LR人脸的判别信息融合到骨干网络中**，**同时保留了原始人脸表示的判别能力。**在IJB-C TinyFace，SCface和QMUL SurvFace数据集上的实验结果证明，与常规的监视人脸识别方法相比，我们的方法具有优越性，同时在常见的识别场景下表现出稳定的性能。

**Introduction**



我们提出了一种新颖的分辨率适应网络（RAN），其中包括用于合成逼真的LR人脸的多分辨率生成对抗网络（MR-GAN）。然后包括特征自适应网络，以逐步学习多分辨率（MR）知识。该框架如图1所示。与[3]不同，它采用GAN生成LR图像作为中间步骤实现图像超分辨率。**我们的MR-GAN旨在直接生成可在大型数据集中进行增强的逼真的LR人脸，并提供先前的多分辨率表示**。生成器采用全局和局部机制来关注不同的区域。在生成器的全局流中，将输入人脸降采样为三个比例，然后传递以提取特定知识。然后，将多分辨率表示逐渐组合并收敛到最低分辨率流中，**以通过空间注意力获得精细的全局人脸**。通过重复连接来自高分辨率子编码器的信息来进行多分辨率融合，并且可以随机选择一种分辨率以细化现实的LR人脸。同时，使用最低尺度人脸的局部区域来获得蒸馏区域，将其与全局人脸聚合以生成逼真的LR人脸。因此，可以使用粗糙但仍具有判别力的人脸来提供低分辨率表示。

遵循生成适应的概念，我们提出了一种新颖的特征自适应网络来指导HR模型融合所生成的LR脸部的判别信息，并保持HR脸部的稳定判别能力。 因此，可以防止通过强制拉近不同域的特征而引起的域移位问题。 具体来说，提出了翻译门以平衡翻译嵌入的来源并逐步保留LR表示。 为了使翻译后的LR嵌入和合成的LR脸部提取的逼真的LR嵌入之间的距离最小化，可以保证HR模型包含足够的LR信息并构造MR嵌入，同时保留人脸细节和轮廓信息。

Contributions：

–我们提出使用多分辨率GAN来合成逼真的LR人脸，这样可以避免低采样人脸产生的伪像。组合不同分辨率的表示，并将其注入最低分辨率的流中以细化LR人脸。全局和局部体系结构都用于生成器和鉴别器中，以增强生成的人脸的真实感。

–我们提出了特征适应网络，**以重定向HR模型**，以便在保留HR表示的同时专注于融合LR信息。**该网络采用翻译门从HR嵌入中逐步提取LR知识，以确保HR模型包含足够的LR信息。**

–我们从IJB-C [29]中选择小脸，并构建名为IJB-C TinyFace的测试集，以利用不受约束的监视人脸识别功能。我们的方法在监视数据集上实现了最先进的性能：SCface [15]，QMUL-SurvFace [9]和IJB-C TinyFace，并在LFW [21]，CALFW [53]，CPLFW [ 52]，AgeDB-30 [30]和CFP-FP [34]。

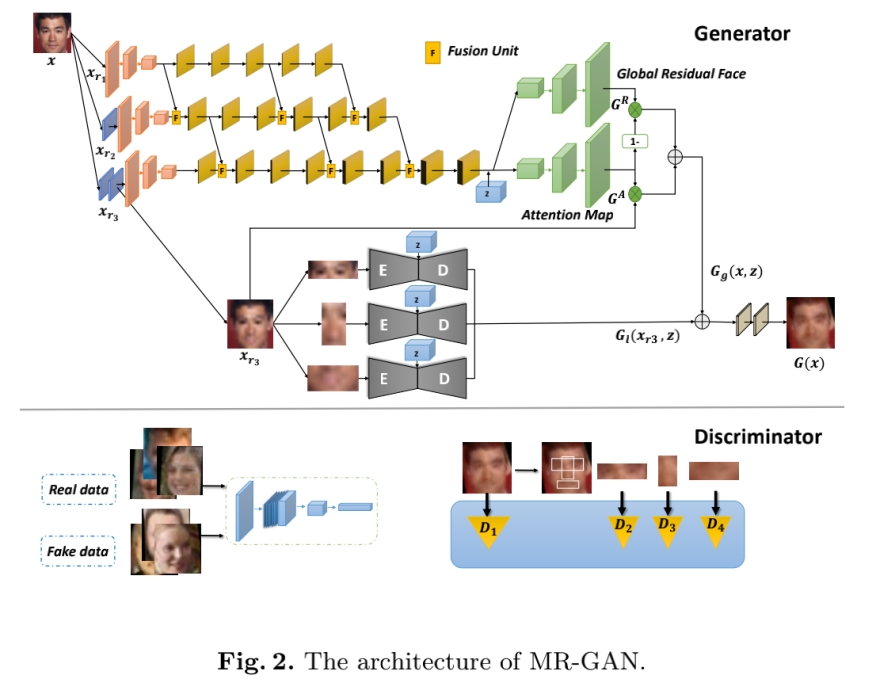
**Method**

无需下采样和双三次线性插值，我们通过MR-GAN获得LR人脸[5，27]，同时避免伪像，从而使我们能够利用未配对的HR人脸，这对于处理无法使用成对的人脸的大规模数据集至关重要。

**采用所提出的自适应网络来提高LR脸部的性能，同时仍保留HR脸部的辨别能力。**如图1所示，我们的方法包括三个步骤：（i）合成逼真的LR面；

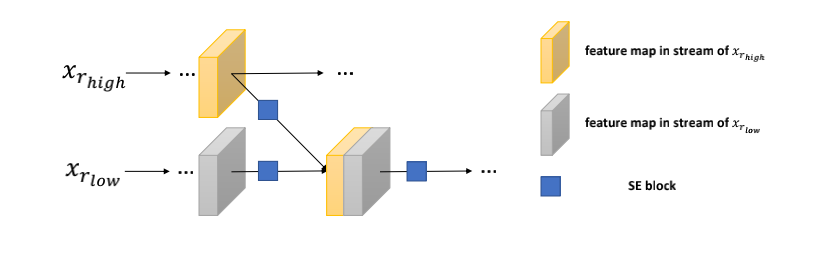
（ii）使用HR人脸和合成的LR人脸作为训练数据集分别训练HR模型和LR模型;

（iii）使用特征自适应网络重新指导HR模型，以学习分辨率稳定的分布。



**Low-Resolution Face Synthesis**

**Resolution-aggregated Generator.**



（图3:说明融合单元如何连接来自不同流的特征图。 在连接之前和之后，由SE块选择表示，并通过更深的通道流入分辨率较低的流中。）

为了最小化HR域和LR域之间的距离，我们首先采用简单的下采样来获得三个模糊度的三个输入：xr1，xr2和xr3，其中xr1保持最高分辨率，而xr3则是最低分辨率。然后，我们使用生成器基于下采样进一步细化全局和局部信息。受到HRNet [36]的启发，我们引入了并行子网，以重复地从更高分辨率的子网中接收信息，并将特征图集成到全局流中。子网采用三个跨步卷积层将人脸编码为特征图。然后使用残差块进一步加深网络，并产生特征图以保持相同的宽度和高度。为了融合来自不同流的信息，采用了融合单元。我们在图3中说明了融合单元的细节，其中所有操作的特征图都是从残差块中学习的。融合单元中的特征图可以表示为{F1r1，F1r2，F1r3 ...，Fkr1，Fkr2，Fkr3}，其中上标k表示第k个残差块的特征图，下标r表示分辨率为r的流的特征图，为了融合不同分辨率的Fr，我们连接了两个特征图以加深通道。例如，可以集成C1×W×H的Far1和C2×W×H的Fbr2以获取（C1 + C2）×W×H的特征图。为了增强融合单元中的分辨率和与身份相关的信息，在特征连接之前和之后整合挤压和激发（SE）块[19]。通过重复的特征融合，将高分辨率的特征图逐渐注入分辨率较低的流中。同时，由于通过连接不同流的特征图可以最大程度地保留多分辨率，因此我们可以注入随机噪声z的向量，以有效地选择和模拟不同程度的分辨率退化。为了解码低分辨率信息并更多地关注与分辨率相关的改变，我们引入了空间注意力以忽略背景。因此，全局流的输出可以总结为：



其中GR（x，z）是输出残差人脸，GA（x，z）表示用于描述对输出Gg（x，z）的贡献的注意力图。因此，重要区域可以由生成器学习，并且无关的像素可以直接从xr3中保留。 局部生成器Gl包含三个相同的子网，它们分别学习以三个中心裁剪的局部块：眼睛，鼻子和嘴巴。 这些区域通过检测到的地标获得并固定。通过传递编码器-解码器流并注入随机向量z，可以细化三个裁剪的局部块，将其与全局人脸Gg（x，z）进一步组合，然后馈入两个1×1跨步卷积层中以生成人脸 G（x）。

**Global-local Focused Discriminator.**

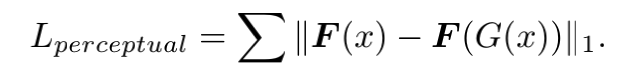
我们采用了一系列判别方法来区分全局和局部区域，从而增强了判别能力。考虑到LR脸部的特征，我们采用与生成器的局部分支相同的接收区域（包括眼睛，鼻子和嘴巴）来构造局部鉴别器，而全局鉴别器则接收整个脸部。如图2所示。这四个鉴别符（Dk，k = 1，2，3，4）注意分别区分不同的区域。与简单的下采样和双三次插值相比，MR-GAN重视保证局部纹理保持固定和自然模糊，并具有很高的视觉质量。

**Loss Function**

我们的MR-GAN的主要目标是生成LR人脸，同时保留身份信息以避免伪像。 提出了几种损失条件来学习逼真的表示。

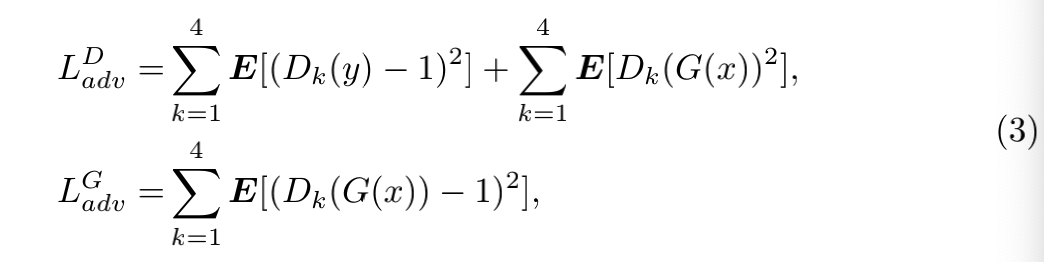
**Perceptual Loss**

为了确保生成的LR人脸保留与输入人脸相同的身份，引入了感知损失以减少高维特征空间中的差异。并通过预训练的专家网络提取高级特征表示F。 因此，损失可以表述为：



**Adversarial Loss**

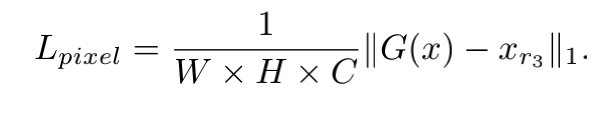
对抗性损失用于从源到目标分布的跨域适应。 损失函数表示如下：



其中x是输入HR面，y表示真实的LR面。下标k指向相应区域的鉴别器。 采用最小二乘损失以确保鉴别器无法区分合成的人脸。

**Pixel Loss.**

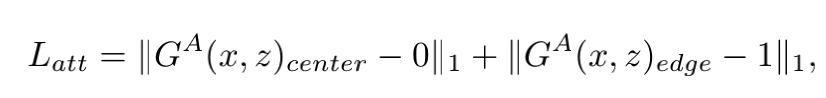
除了专门设计的对抗性判别和身份惩罚外，还采用了图像空间中的L1损失来进一步完善简单的下采样并弥合输入输出差距，其定义如下：



如前所示，xr3是最低分辨率的输入，可用于加快收敛速度和稳定优化。

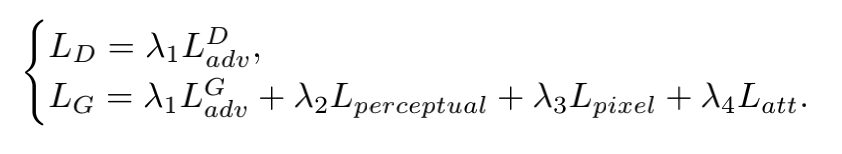
**Attention Activation Loss.**

如公式5所示，当GA（x）中的所有元素都饱和为0时，所有输出都被视为全局输出。 为了防止学习与身份无关的信息，采用注意力激活损失来将激活限制在重要mask上，而忽略背景周围的信息。 因此损失函数可以写成：



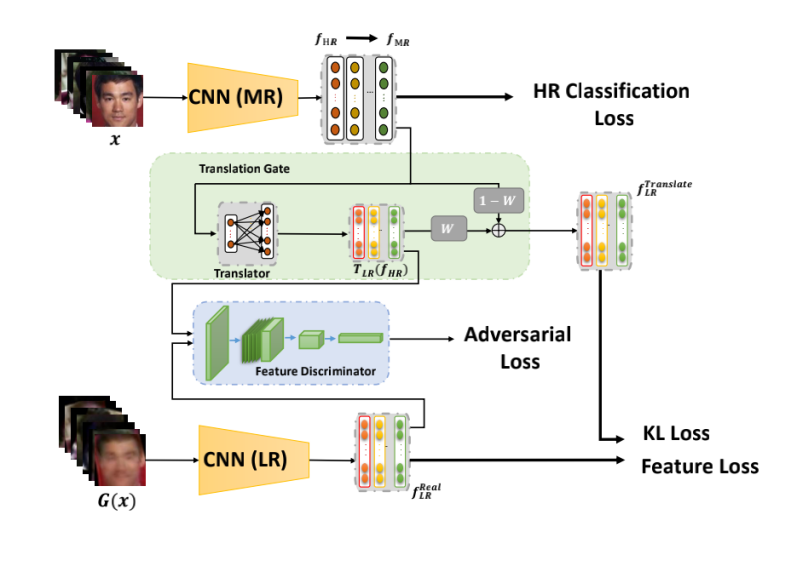
其中GA（x，z）center表示注意力图的85×82中心patch，GA（x，z）edge是注意力图的边缘。

总之，我们有四个用于生成LR脸的损失函数，并使用超参数λ1，λ2，λ3和λ4进行平衡。 总体目标是：

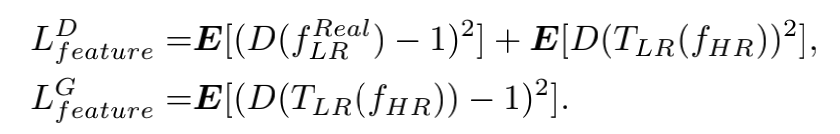


**Feature Adaption Network**

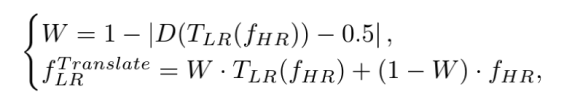
由于大型数据集中缺少足够的LR人脸，我们建议添加生成的目标样本以平衡多分辨率表示。 然而，由于HR域和LR域之间的**域移位**，很难将具有相同身份的简单最小化距离的方法直接应用于监控环境下的人脸识别。为了克服这个问题，我们提出了**一种特征自适应网络来保留HR域的判别能力**，并将其应用于**动态提高LR域的竞争能力**。



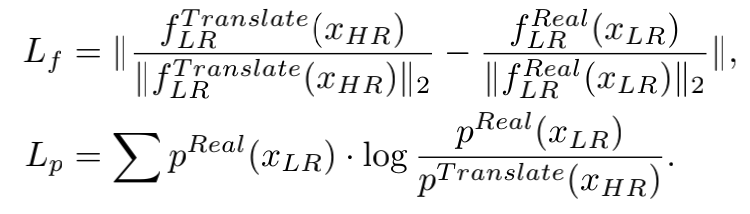
整个框架如图4所示，其中包含两个流。底部的流由生成的LR人脸训练，以提供逼真的LR表示，并在随后的适应学习中进行固定。顶部的Stream用于学习最终的多分辨率嵌入fMR。为了保持HR人脸的判别性，我们采用ArcFace [10]作为分类损失LHR c，使顶流模型直接由HR人脸监督。同时，为了提高在LR人脸上的性能并**通过直接最小化域gap**来避免HR特征空间恶化，我们提出了**翻译门**。翻译门**采用翻译器来平衡fHR的LR分量并确定ftranslate LR的源头。**转换器包括两个批处理归一化层，分别**是ReLU和完全连接的层**，在放大L​​R表示以获得LR特征TLR（fHR）方面起中间作用，使HR特征fHR专注于保存LR信息。通过逐渐翻译真实的LR特征，流顶部的HR模型可以保留更多的LR表示，以获得多分辨率嵌入fMR。为了实现此目标，我们应用了低分辨率对抗网络，以确保翻译后的LR嵌入TLR（fHR）足够真实，以至于使鉴别器产生混淆。采用LSGAN [28]将它们组合在一起。损失函数如下：



通过采用LSGAN，| D（TLR（fHR））- 0.5 | 用于表示已翻译的LR特征的置信度。 **鉴别器的输出越接近0.5**，则翻译的LR特征越真实，这表示f HR可以保留更多的LR信息并**获得具有平衡多分辨率知识的fMR**。随着置信度的提高，fHR还可以保存并提供足够的LR表示，而无需翻译。 因此，我们的翻译门采用加权架构来确定最终的LR特征：



其中W是平衡TLR（fHR）和fHR的权重。在获得f Translate LR之后，我们将L1损失和KL损失相加，以学习特征和概率表示的低分辨率人脸分布，从而进一步将翻译后的嵌入逼近真实的嵌入。 损失如下：



考虑到fHR在训练的早期包含有限的LR表示，TLR（fHR）在特征和概率监督中起主导作用。 然后，随着HR特征可以逐渐保留并提供更逼真的LR表示，W将保持在稳定范围内以平衡两个低分辨率知识来源。通过这种加权翻译，**fHR可以保留足够的LR表示，以构建分辨率鲁棒的嵌入**。因此，总损失可以看作是：



**Experiments**

**Experiment Settings**

在本节中，我们介绍了提出的分辨率自适应网络的结果。CASIA-WebFace [42]用作HR脸部来训练MR-GAN和特征适应网络。

**真实的LR人脸选自MillionCelebs [50]**。我们使用MTCNN [48]进行人脸检测和对齐。所检测到的landmarks被用来测量眼睛的中心点与嘴巴中心之间的距离。距离小于30且大于10的人脸被选择为逼真的LR人脸。为了评估特征自适应网络的性能，

我们利用34层Resnet[17]作为骨干，并采用SCface [15]，QMUL-SurvFace [9]和IJB-C的低分辨率子集[29]（IJB-C TinyFace）作为测试集。 IJB-C [29]是一个基于视频的人脸数据库，其中包含自然的分辨率变化。我们遵循相同的规则来选择逼真的LR人脸。采用所有检测到的LR人脸，并为每个锚点选择具有相同标识的人脸，以构造正对，包括158,338个真实比较。遵循IJB-C 1：1协变量验证协议，在IJB-C TinyFace中使用相同的39,584,639个负对。 SCface [15]包含130个对象的人脸图像。 [27]之后，有80个主题用于测试，其他50个主题用于微调。在使用HR人脸作为画廊集并分别以4.2m（d1），2.6m（d2）和1.0m（d3）拍摄的LR图像的情况下进行人脸识别。 QMUL-SurvFace [9]由在监视摄像机下捕获的超低分辨率人脸图像组成。

**Implementation Details**

所有的训练和测试面孔都被**裁剪并对齐为112×112**。在 MRGAN中，我们通过使用Adam优化迭代迭代地最小化鉴别器和发生器损失函数来训练鉴别器和发生器。每5个生成器迭代使用一次像素批判。 **MobileFaceNets [6]被用作专家网络**，并且所有参数都是固定的。根据经验将超参数设置如下：λ1= 2，λ2= 20，λ3= 20，λ4= 0.4并批量size =16。我们将优化器的超参数设置为：β1= 0.5，β2= 0.999，学习率= 0.0002。特征适应网络中采用ArcFace作为分类损失。根据[10]，特征比例和角裕度m分别设置为64和0.5。我们将批次大小设置为256，以训练预训练的HR和LR模型。获取MR嵌入需要三个步骤。**首先，我们使用CASIA-WebFace对ResNet-34进行预训练，以获取HR模型**。学习率从0.1开始，以60,000,100,000和140,000次迭代除以10。其次，**我们采用生成的LR-CASIA-WebFace作为训练集来微调HR模型，以获得LR模型**。学习率从0.01开始，并在50,000和100,000次迭代中除以10。为了模拟更多分辨率变化，训练LR模型时会添加随机高斯模糊。**最后，继续通过使用带有固定LR模型的间接监督的HR人脸来微调HR模型以训练MR模型。**

在此步骤中，批量大小设置为200，学习率从0.01开始，在50,000次迭代时，学习速度除以10。可以将超参数设置如下：α= 0.05，β= 0.04，γ=10。我们采用SGD优化进行识别，而采用Adam优化进行对抗学习。 LG功能已更新并在以下位置使用

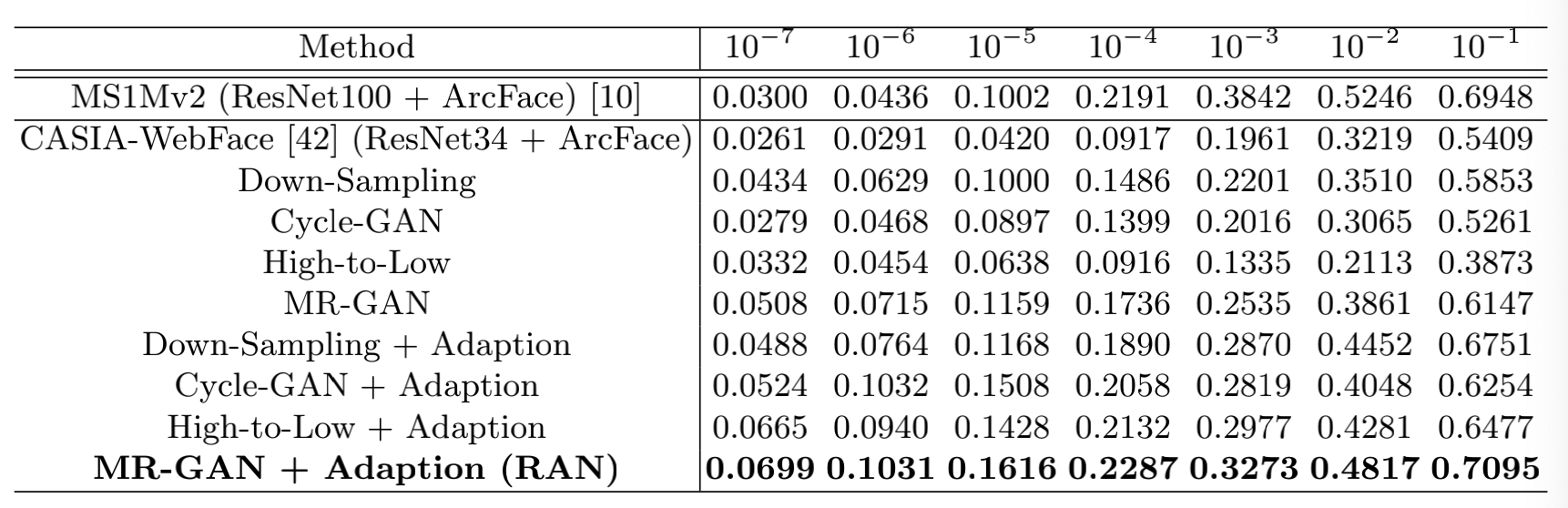
每4个鉴别器迭代。有关网络体系结构的完整详细信息，请参阅补充材料。

**Ablation Study**

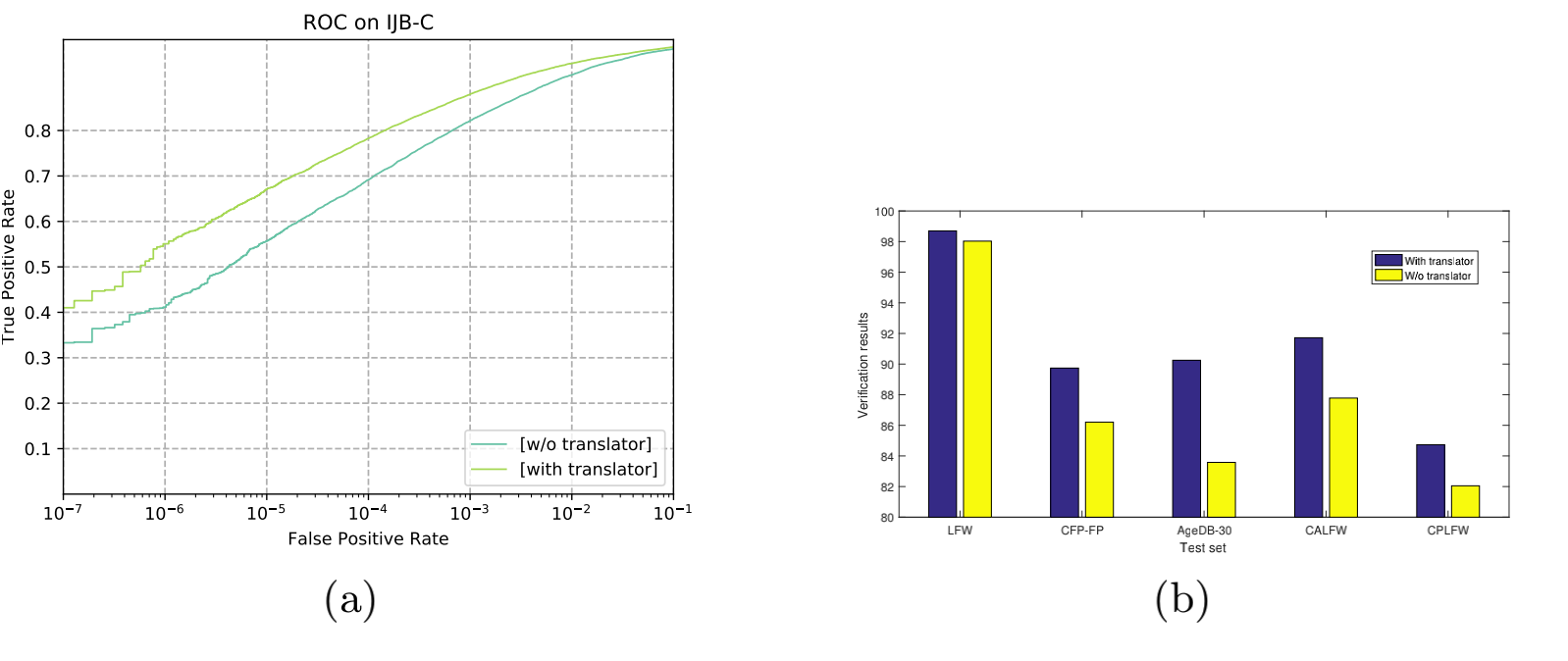
**Effffects of LR Face Synthesis.**

由于现有的大规模数据集（例如CASIA-Webface [42]和MS-Celeb-1M [16]）包含许多HR面孔，因此我们的方法旨在生成具有不同分辨率的LR面孔以增强训练集。但是，大多数现有作品都采用下采样来获取LR人脸，这与实际环境不符。如图5所示，由**下采样产生的人脸充满了不规则的扭曲和噪声**。当分辨率降低时，基于GAN的合成方法可以保持人脸的真实感。但是，Cycle-GAN [55]生成的脸部过于光滑。 Bulat等。 [3]旨在采用从高到低和从低到高的人脸超分辨率。他们忽略了保留脸部细节的信息，并在LR脸部进行了有限的监督。因此，由高到低生成器生成的LR人脸不能直接用于识别。相比之下，我们的MR-GAN可以集成多分辨率信息以利用特定的表示形式，并且可以将更多精力集中在局部区域的模拟上，从而获得粗糙但具有区别性的细节。在补充材料中可以找到更多的可视化内容。

为了与人脸识别结果进行定量比较，**我们在IJB-C TinyFace上评估了不同的方法，并在表1中报告了结果。**我们将CASIA-WebFace的所有面孔转换为LR面孔，包括：下采样，GAN循环[ 55]，从高到低[3]和 MRGAN，并采用生成的训练集来微调HR模型。 结果从第3行到第6行进行了描述。通过适应，性能得到了进一步提高。 由于从高到低[3]生成的脸很小，无法直接识别，因此结果相对较低。 但是，从高到低在适应学习过程中仍然提供了足够粗糙的细节，这显示了有效性。 为了更好地证明RAN的效果，我们使用更大的数据集和更多的参数报告模型[10]的结果，该结果显示在顶部。 我们的方法利用较小的模型和训练集来达到相同的性能，甚至在某些情况下甚至超出了它们。



**Effffects of MR Feature Adaption.**



为了防止由于域gap而直接最小化HR和LR域的距离，建议翻译门使用翻译器来平衡翻译后LR特征的来源。 如果不使用转换器，则直接采用fHR来最小化不同域之间的距离。 在图6（a）中，通过直接最小化特征和概率分布的距离，辨别能力随着FAR的降低而迅速下降。 在图6（b）中，LFW的精度降低到97.7。 然而，在翻译器的中间作用下，翻译门可以采用加权架构逐步生成TLR（fHR）。 因此，LFW的精度可以保持在98.7。 在IJB-C和高分辨率测试集上保存的结果表明，我们的带翻译门的MR嵌入可以应用于两个域，并且在处理困难情况方面显示出显着的有效性。

**Performance on Large-scale Dataset.**

为了显示我们的RAN在大规模数据集上的有效性，将包含5,084,598张面孔和97,099位受试者的清洁过的MS-Celeb-1M [16]用作训练集。 ResNet-50和ArcFace被用作基本训练体系结构和损失函数。 本实验采用相同的训练步骤。 HR，LR和MR模型的结果如表2所示。由于大规模数据集已经包含许多低分辨率图像，因此仅采用ArcFace损失进行监督才能在HR模型中获得较高的性能。 通过使用MR-GAN将所有数据集转换为LR数据集，LR模型优于HR模型（FAR小于10 4）。 此外，我们的RAN通过集成多分辨率知识在所有情况下均实现了最高性能。

**Compare with SOTA Methods**

**Comparisons on SCface.**

SCface定义了人脸识别协议。对于每个被摄对象，监视相机在三个距离处（每个距离五张脸）拍摄了15张脸，而数码相机则拍摄了一张正面的人脸照片。为了公平比较，**我们将SOTA人脸识别方法ArcFace [10]作为HR建模并遵循[27]在SCface上进行微调**。比较的方法更多地集中在最小化不同分辨率下的类内距离，但是这些方法直接忽略了类间距离，而忽略了分辨率差距。而且他们只是采用下采样来增加分辨率的多样性，并提供与实际情况不符的成对的多分辨率人脸。选择性知识提炼[13]采用HR模型作为教师，并采用LR模型作为学生，以恢复LR模型区分人脸细节的能力。由于高分辨率信息已经丢失，因此无法恢复足够的表示。相反，我们的RAN致力于通过分辨率适应来保留来自HR功能的LR信息，从而可以学习足够的多分辨率知识并获得最佳性能。

**Comparisons on QMUL-SurvFace.**

QMUL-SurvFace包含非常低的LR面孔，这些面孔是从真实的监控视频中提取的。 我们将RAN与人脸超分辨率（SR）方法和常见识别方法进行比较。 如表4所示，我们进行人脸识别。 大的边际损失（ArcFace和SphereFace）已在大规模数据集中实现了SOTA结果。 因此，它们提高了HR域的性能，也可以应用于LR域。 然而，这些人脸SR方法难以恢复身份信息并且更多地关注视觉质量，从而不可避免地降低了性能。 通过从HR面孔动态提取特征空间中的MR知识，我们与脸部SR和常见识别方法相比，该方法的性能更好。

**Conclusion**

本文提出了用于现实的LR人脸合成和监视人脸识别的分辨率自适应网络（RAN）。 我们旨在生成用于数据增强的LR人脸，并弥合交叉分辨率差距。 在RAN中，MR-GAN采用多分辨率和全局局部架构，以随机分辨率模糊人脸，以生成身份保留的逼真的LR人脸。 为了使用LR人脸更好地与LR人脸和HR人脸匹配，提出了特征自适应网络来增强LR知识并逐步平衡多分辨率表示。 SOTA结果可用于监视人脸识别。