1. **pre-knowledge**

Hilbert空间：完备的内积空间，是有限维欧氏空间的推广。

Layer-layer comparison **=>** similarity between networks (Require same dimension)



pearson系数只能刻画线性相关度，对于非线性无能为力，但是distance correlation可以

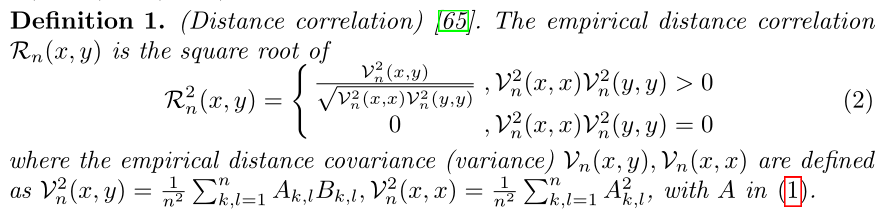
For different layers and dimensions:

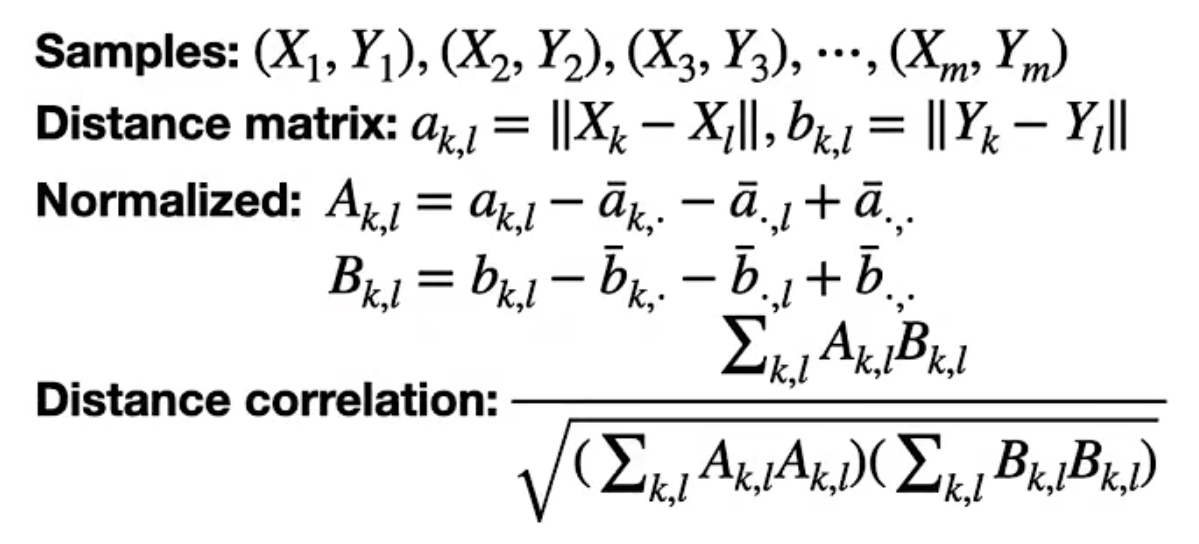
Canonical Correlation Analysis (CCA)

(but the estimation of projection matrices during training is problematic)

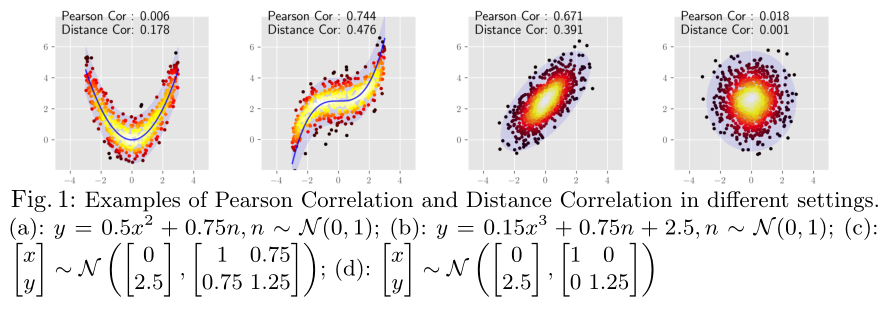
Resolution: partial distance correlation

Covariance 协方差



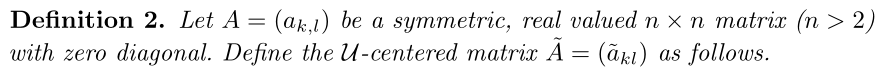


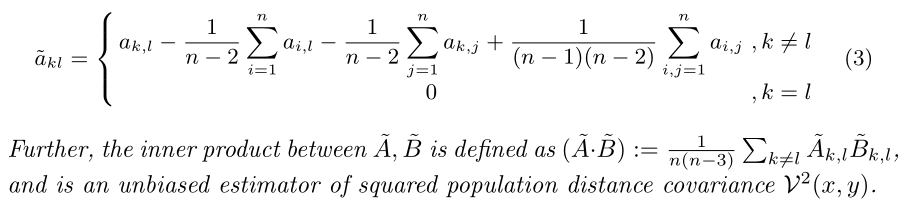
Examples:



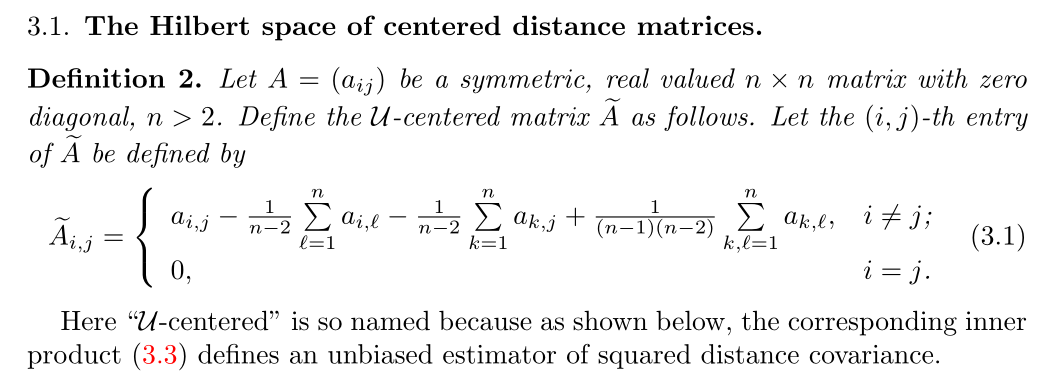
第2、3张图的解释？？（只是因为有线性的趋势，导致P-Cor 优于 D-Cor？？）

inner product 内积

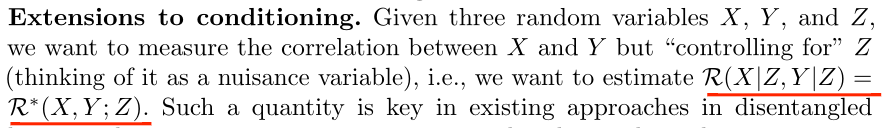


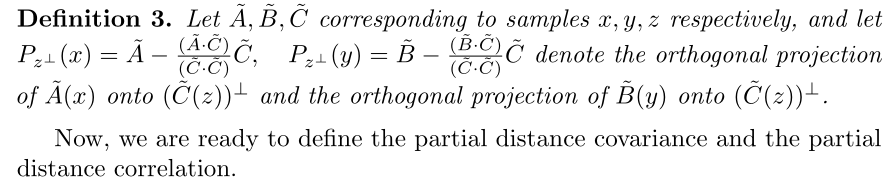


U-centered matrix 来源于该文章：**“PARTIAL DISTANCE CORRELATION WITH METHODS FOR DISSIMILARITIES“**

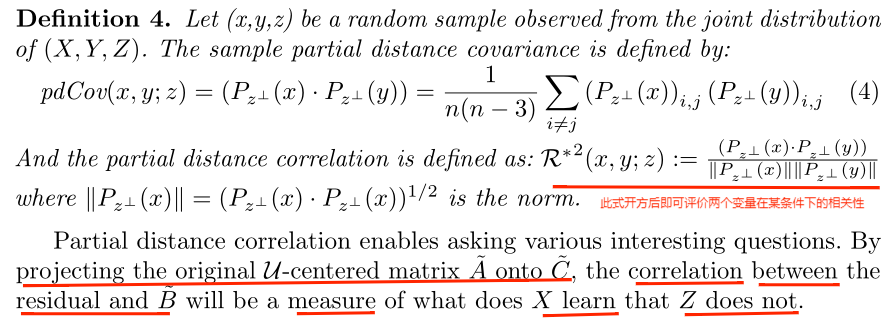
****

unbiased estimator 无偏估计量





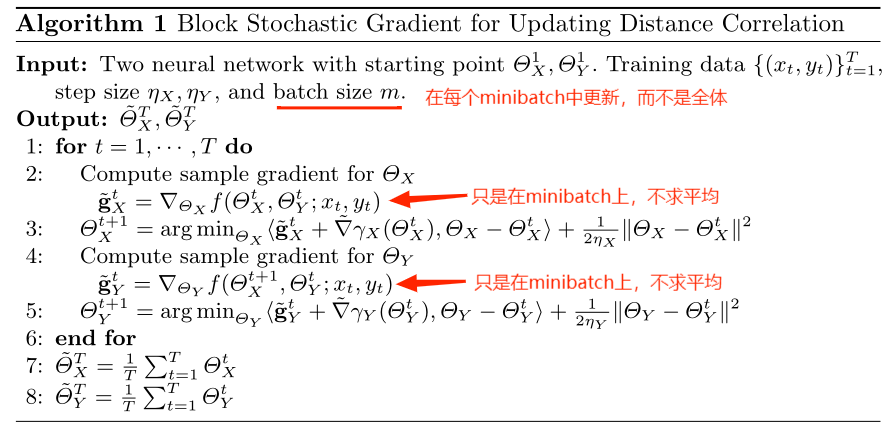
orthogonal projection 正交投影



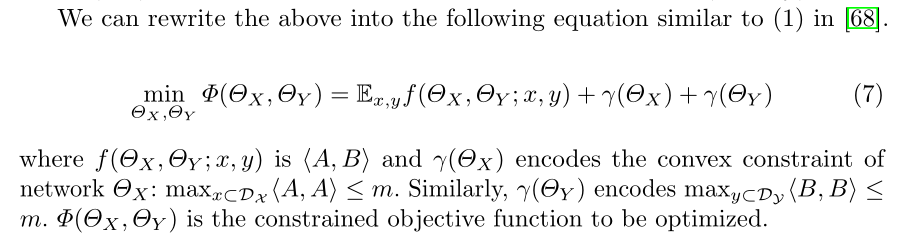
详细证明见该文章：**“PARTIAL DISTANCE CORRELATION WITH METHODS FOR DISSIMILARITIES“**

**二、Optimizing Distance Correlation in Neural Networks**

**Block stochastic gradient iteration：**

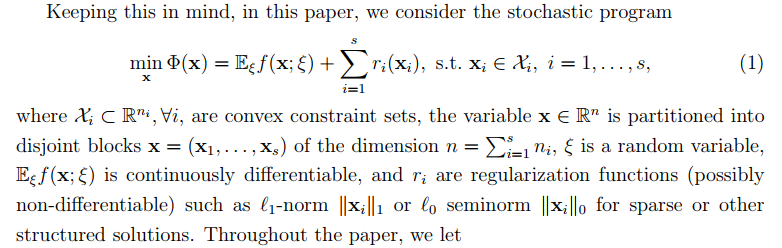




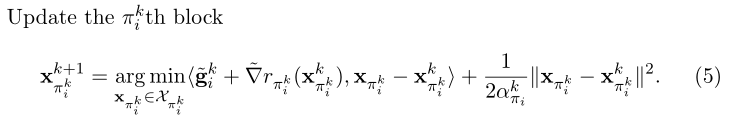


每个网络有一个正则化项（凸约束，≤m），两个网络相关性尽可能低

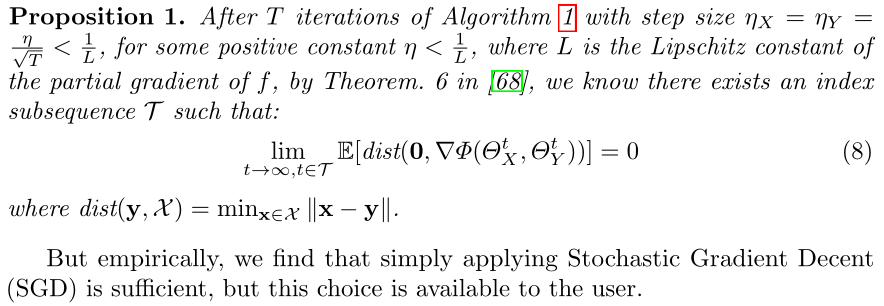
**（1）in [68]：**



**（5）in [68]：**



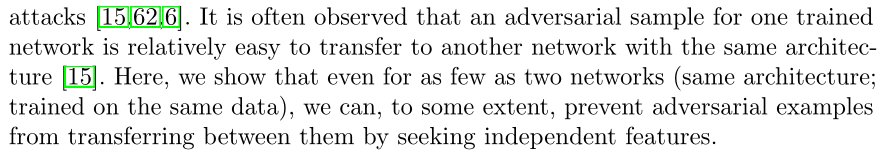
**Convergence analysis**



**三、Use Cases**

1. **diverge training:** 确保多个子网络的独立性

improve robustness (train several networks and resemble)



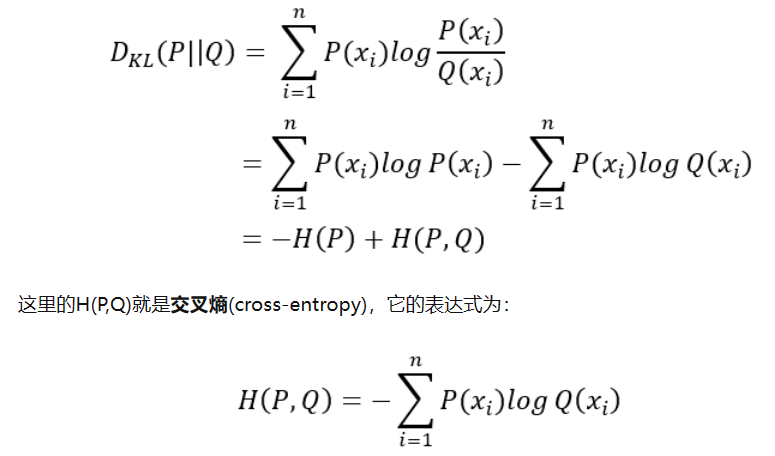
f1(x)只用LossCE训练，f2(x)用Losstotal训练





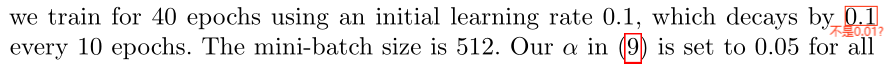
α：常数

LossCE：（cross-entropy loss，交叉熵损失）：



(DC to impose each subnetwork to be independent **=>** more diverse distribution of networks) (一部分样本被攻击，剩下样本因相互独立，仍可以不受影响) 证明？

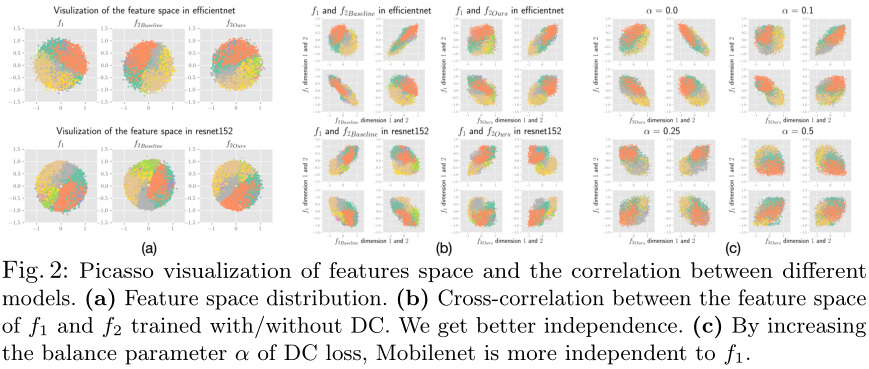
|  |  |  |
| --- | --- | --- |
| Network | Dataset | Details |
| Resnet 18 | CIFAR10 | SGD: 0.9 ; initial learning rate: 0.1 ; epochs: 200 ; minibatch: 128 ; α: 0.05 |
| mobilenet-v3-small | ImageNet | Initial learning rate: 0.1 (for 40 epochs) (decay 0.1 every 10 epochs) ; α: 0.05 |
| efficientnet-B0 |
| Resnet 34 |
| Resnet152 |



attack methods: FGM(fast gradient sign method), PGD(projected gradient descent

method) (strongest attack) ; scale ϵ of the adversarial perturbation is

chosen from {0.03, 0.05, 0.1} ; max\_iteration of PGD is 40



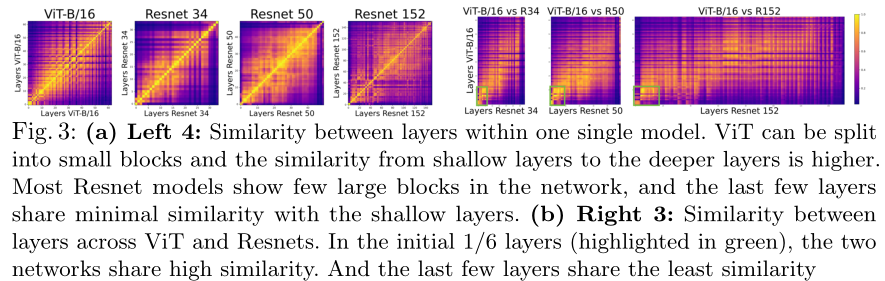
Result: DC can train more independent network

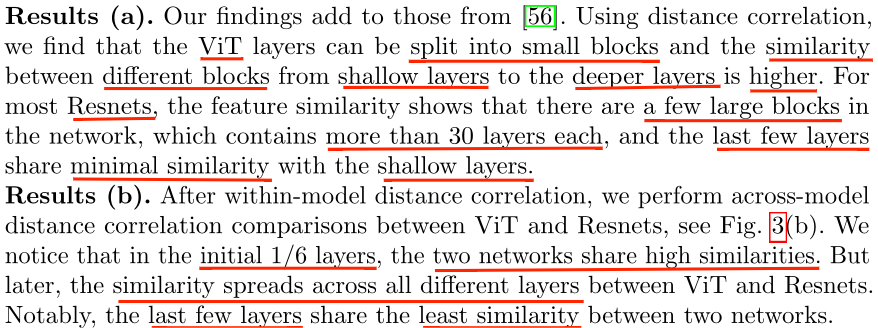
1. **network conditioning:**

Network: Vit, Resnet

Goal: demonstrate DC is a reasonable alternative for CKA(Centered Kernel Alignment) in these settings

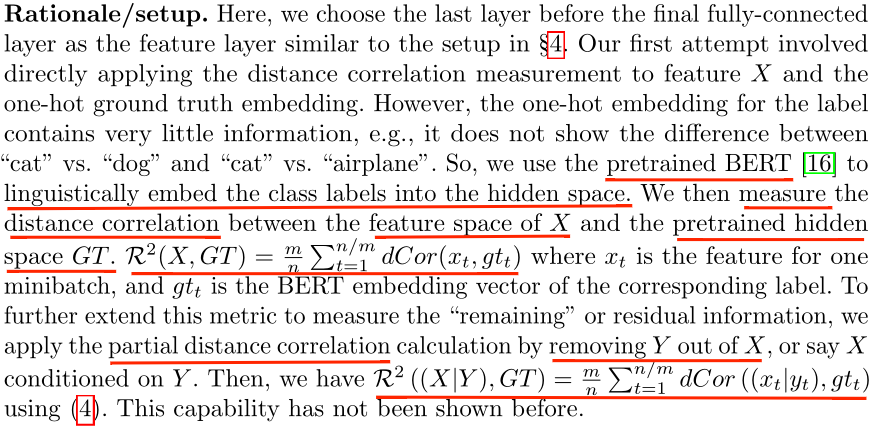
Settings: Vit(embedding layer and all the normalization, attention, and fully connected layers, total layer-number is 63) ; Resnet(all convolutional layers and the last fully connected layer)





**quantitatively show the difference：**

**(measure the similarity between the features of the network X and the gt. Labels)**

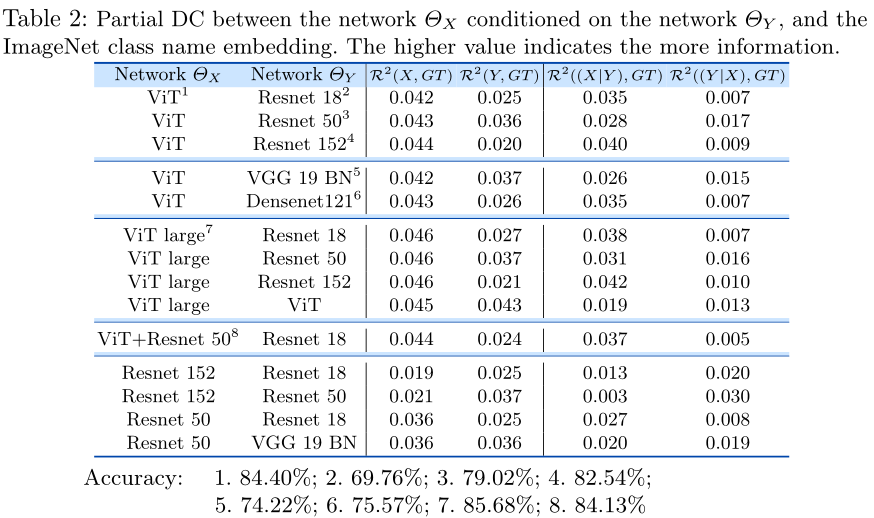


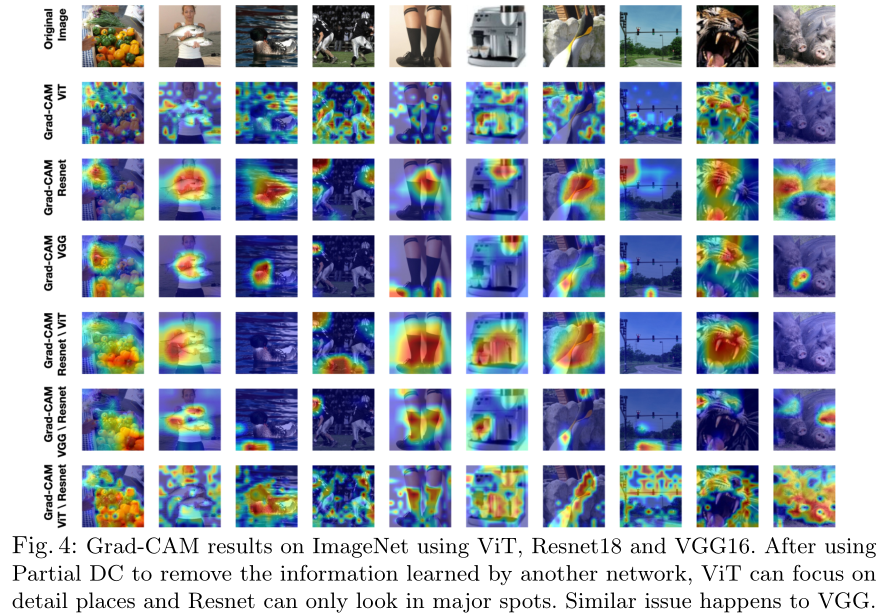
**To maximize partial distance correlation**

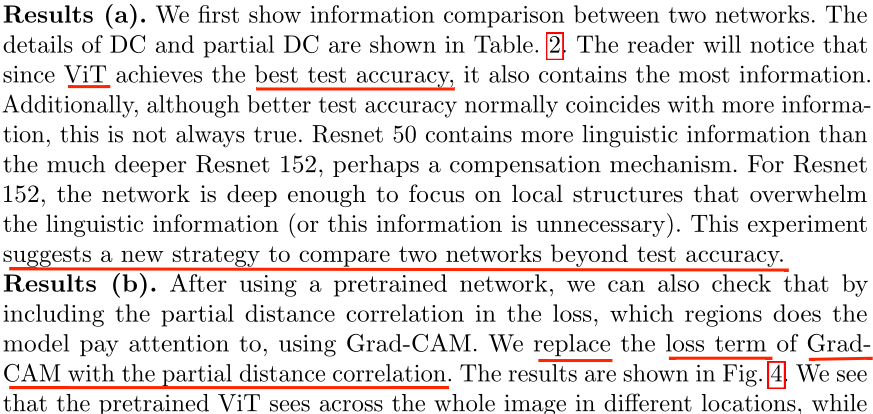


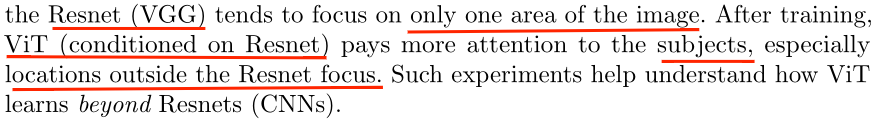
Learning rate: 1e-5

α: 1









Result: remove resnet from vit, but still can learn meaningful result

1. **disentanglement:**

for encoder decoder model: can ask one node in the latent space (represents the attribute of interest) and another node (represents the residual information) should be independent ——entanglement

**Key: to make the factors in the latent variables independent**

In [21], one divides the latent variables into two categories:

(a) attributes of interest – a set of semantic and interpretable attributes, e.g. hair color and age;

(b) residual attributes – the remaining information

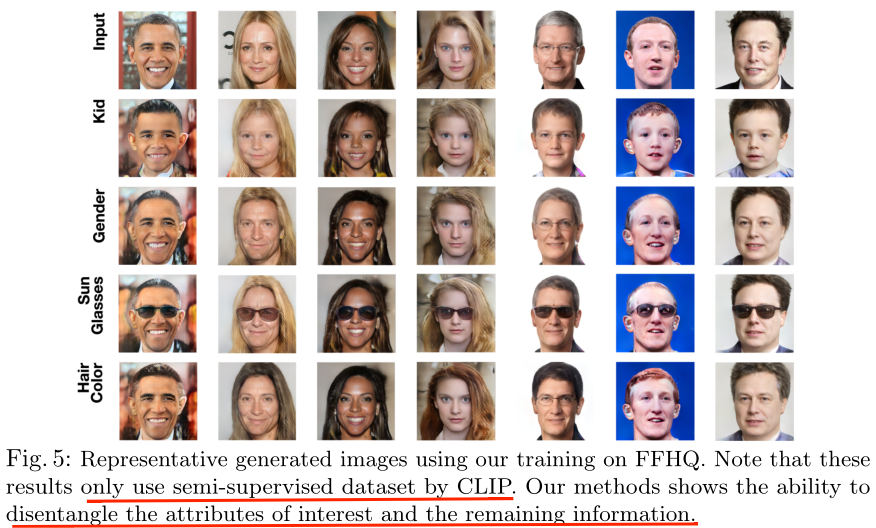


**替换为**

Dateset: FFHQ

generator architecture: StyleGAN2

training dataset: use CLIP to partially label the attributes to generate the semi supervised dataset



Result: change several attributes of interest by using semi-supervised label

**四、Idea**

使用PDC or DC：

1. 进行特征解耦，独立化网络学习到的特征，便于分块解释学习过程。
2. 不同网络间的相似度比较，或许可以构造出基网络（最简单的神经网络，学习过程可解释）然后通过不断迭代相似度比较，可解释较复杂的神经网络？