'Method' Analysis

Written by: 류채은 Chaeeun Ryu (superbunny38 *at* gmail *dot* com)

<Understanding the technologies and math terms used in the paper>

Bringing Old Photos Back to Life Paper

Supplementary Material

Terminology

- latent space: refers to an *abstract multi-dimensional space* containing feature values that we cannot interpret directly, but which encodes a meaningful internal representation of externally observed events.
- LSGAN: Least Squares Generative Adversarial Networks VAE(Variational AutoEncoder)
- VGG:

> https://s3-us-west-2.amazonaws.com/secure.notion-static.com/9b2d47a3-cc6a-40d6-834c-a061d473320b/VGG.pdf

Problems:

1. Generalization issue

Old photos contain far more complex degradation that is hard to be modeled realistically and there always exists a **domain gap between synthetic and real photos**. As such, the network usually cannot generalize well to real photos by purely learning from synthetic data.

2. Mixed degradation issue

The defects of old photos are a compound of multiple degradations, thus essentially requiring different strategies for restoration.

- Unstructured defects
    - film noise, blurriness and color fading, etc.
    - can be restored with spatially homogeneous filters by making use of surrounding pixels within the local patch
- Structured defects
    - scratches and blotches
    - should be inpainted by considering the global context to ensure the structural consistency

Method #1: Restoration via latent space translation

- usage of VAE ❣️ VAE (Variational Autoencoder)

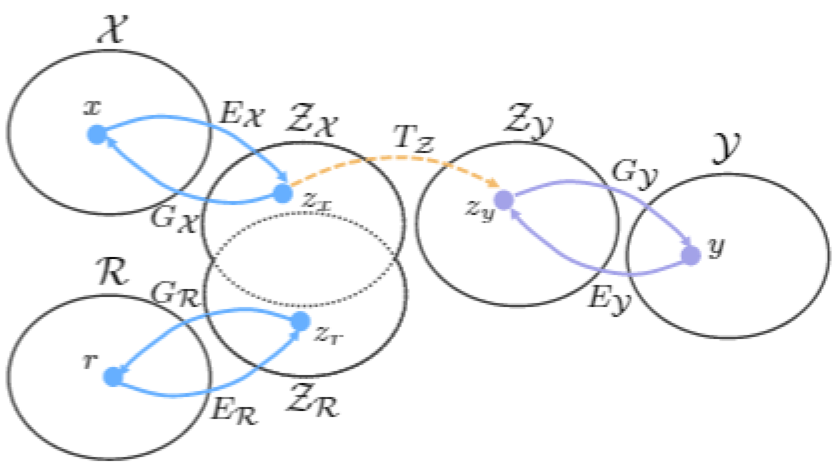


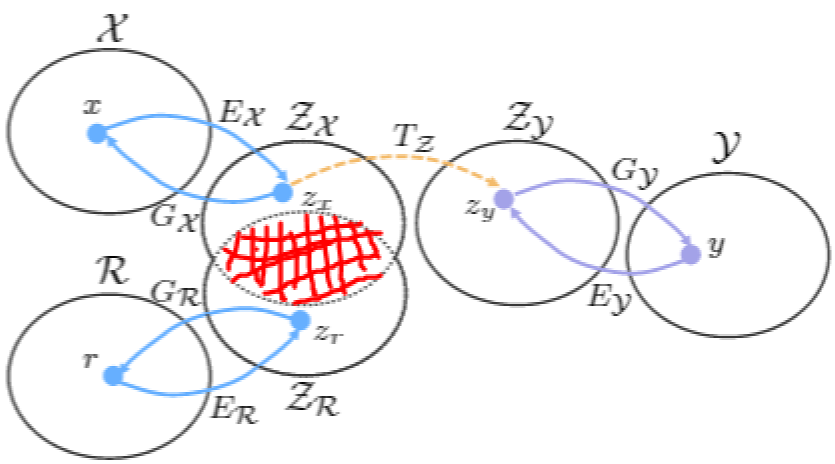
Illustration of translation method with three domains

- image translation problem
- translate images across **three domains**
  1. R : the real photo domain
  2. X : the synthetic domain; where images suffer from artificial degradation
  3. Y : ground truth domain; where comprises images without degradation and corresponding to X
- Images: r∈ R, x∈ X, y∈ Y
  - x and y are paired by data synthesizing,i.e., x is degraded from y

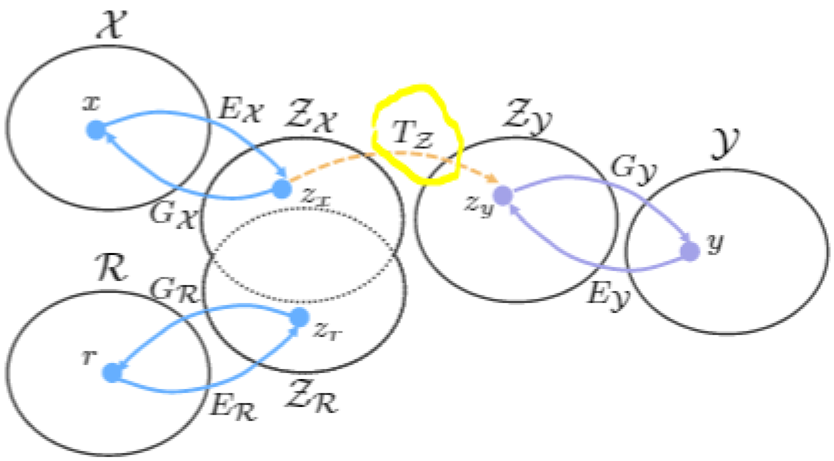
Process

Step #1. we propose to map R, X, Y to corresponding latent spaces via

- $E_R : R \to Z_R$
- $E_X : X \to Z_X$
- $E_Y : Y \to Z_Y$



$Z_R \approx Z_X$: we align latent spaces of **synthetic images and real old photos into the shared domain** by enforcing some constraints because both are corrupted; sharing similar appearances. This aligned latent space encodes features for all the corrupted images, either synthetic or real ones.

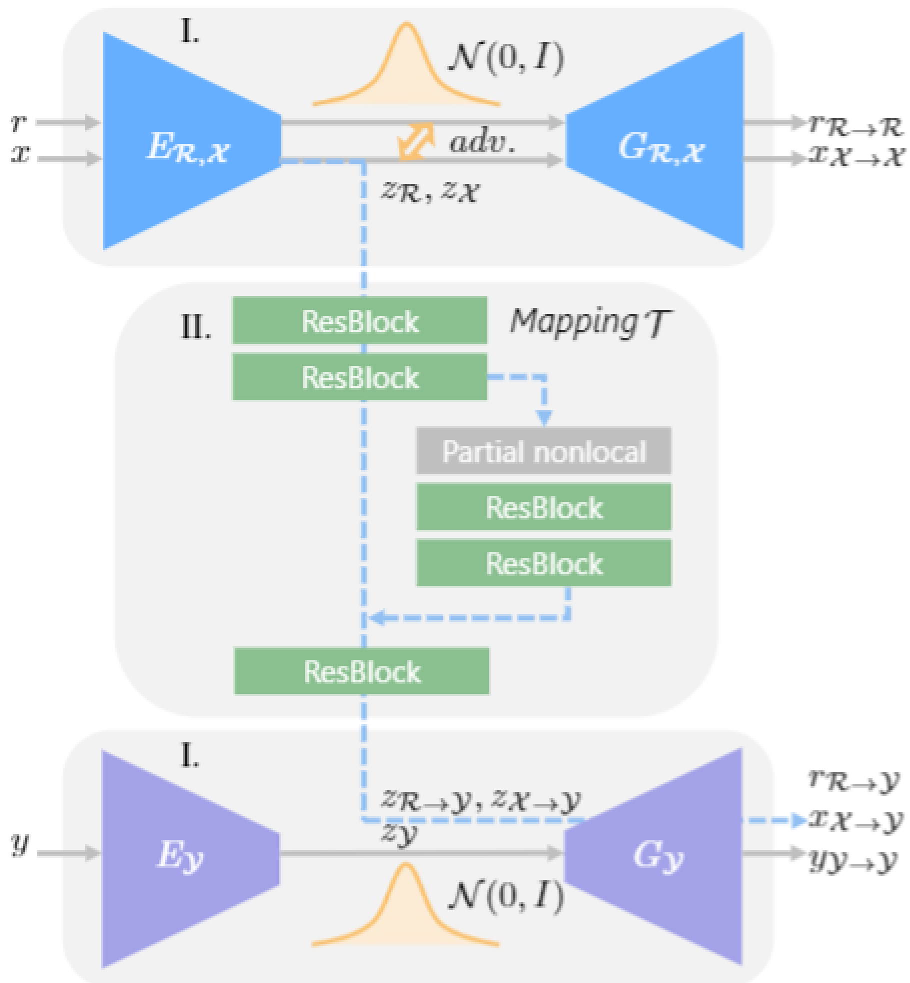


T_Z $=Z_X \rightarrow Z_Y$: we learn the translation from the latent space of **corrupted images,** Z_X **, to the latent space of ground truth,** Z_Y

Z_Y can be further reversed to Y through generator $G_Y : Z_Y \rightarrow Y$.

<u>**Final restoration formula of latent space translation:**</u>

$r_{R \rightarrow Y} = G_Y \circ T_Z \circ E_R(r)$



Architecture of restoration network

I. Domain alignment in the VAE latent space (I.)

- Assumption: R and X are encoded into the same latent space.
- Concept: Utilize variational autoencoder (VAE) to encode images with compact representation, whose domain gap is further examined by an adversarial discriminator. ❣️ GAN(Generative Adversarial Network)
- Process:
  - 1st stage:
    - VAE_1:

- Old photos {r} & synthetic images {x}, encoder $E_{R,X}$ and generator $G_{R,X}$
- Premise: **images from both corrupted domains(x,r) can be mapped to a shared latent space.**
- their domain gap(x,r) is closed by jointly training an adversarial discriminator
- optimization:
  - 💙Objective with r

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}_1}(r) = \ & \mathrm{KL}(E_{\mathcal{R},\mathcal{X}}(z_r|r)||\mathcal{N}(0,I)) \\ & + \alpha\mathbb{E}_{z_r\sim E_{\mathcal{R},\mathcal{X}}(z_r|r)}\left[\|G_{\mathcal{R},\mathcal{X}}(r_{\mathcal{R}\to\mathcal{R}}|z_r) - r\|_1\right] \\ & + \mathcal{L}_{\mathrm{VAE}_1,\mathrm{GAN}}(r)\end{aligned}$$

  - 💙differentiates Z_R, Z_X, loss:

$$\begin{aligned}\mathcal{L}^{\mathrm{latent}}_{\mathrm{VAE}_1,\mathrm{GAN}}(r,x) = & \ \mathbb{E}_{x\sim\mathcal{X}}[D_{\mathcal{R},\mathcal{X}}(E_{\mathcal{R},\mathcal{X}}(x))^2] \\ & + \mathbb{E}_{r\sim\mathcal{R}}[(1 - D_{\mathcal{R},\mathcal{X}}(E_{\mathcal{R},\mathcal{X}}(r)))^2].\end{aligned}$$

  - 💙total objective function for VAE_1:

$$\min_{E_{\mathcal{R},\mathcal{X}},G_{\mathcal{R},\mathcal{X}}} \max_{D_{\mathcal{R},\mathcal{X}}} \mathcal{L}_{\mathrm{VAE}_1}(r) + \mathcal{L}_{\mathrm{VAE}_1}(x) + \mathcal{L}^{\mathrm{latent}}_{\mathrm{VAE}_1,\mathrm{GAN}}(r,x).$$

- VAE_2:
  - ground true images {y}, the encoder-generator pair{E_Y, G_Y}
  - trained for clean images

- 2nd stage:
    - With VAEs, images are transformed to compact latent space

II. Restoration through latent mapping (II.)

- learn the mapping that restores the corrupted images to clean ones in the latent space
- leverage the synthetic image pairs x, y and propose to learn the image restoration by mapping their latent space (the mapping network M)
- 3 Benefits:
    1. As R and X are aligned into the same latent space, the mapping from Z_X to Z_Y will also generalize well to restoring the images in R
    2. the mapping in a compact low-dimensional latent space(code in VAE) is in principle much easier to learn than in the high-dimensional image space
    3. The generator G_Y can always get an absolutely clean image without degradation given the latent code z_Y mapped from Z_X, whereas degradations will likely remain if we learn the translation in pixel level
- Process:
    1. Get $r_{R \to Y}$, $x_{X \to Y}$ and $y_{Y \to Y}$ be the final translation out-puts for r,x and y, respectively.
    2. solely train the parameters of the latent mapping network T and fix the two VAEs
        - Loss function L_T (imposed at both the latent space and the end of generator G_Y)

$$\mathcal{L}_{\mathcal{T}}(x, y) = \lambda_1 \mathcal{L}_{\mathcal{T},\ell_1} + \mathcal{L}_{\mathcal{T},\text{GAN}} + \lambda_2 \mathcal{L}_{\text{FM}}$$

L_T consists of three terms:

a. the latent space loss

(penalizes the l_1 distance of the corresponding latent codes)

$$\mathcal{L}_{\mathcal{T},\ell_1} = \mathbb{E}\,\|\mathcal{T}(z_x) - z_y)\|_1$$

b. the adversarial loss

GAN(Generative Adversarial Networks)

(still in the form of LSGAN, to encourage the ultimate translated synthetic image $x_{X \to Y}$ to look real)

$$\mathcal{L}_{\mathcal{T},\mathrm{GAN}}$$

c. feature matching loss

(to stabilize the GAN training)

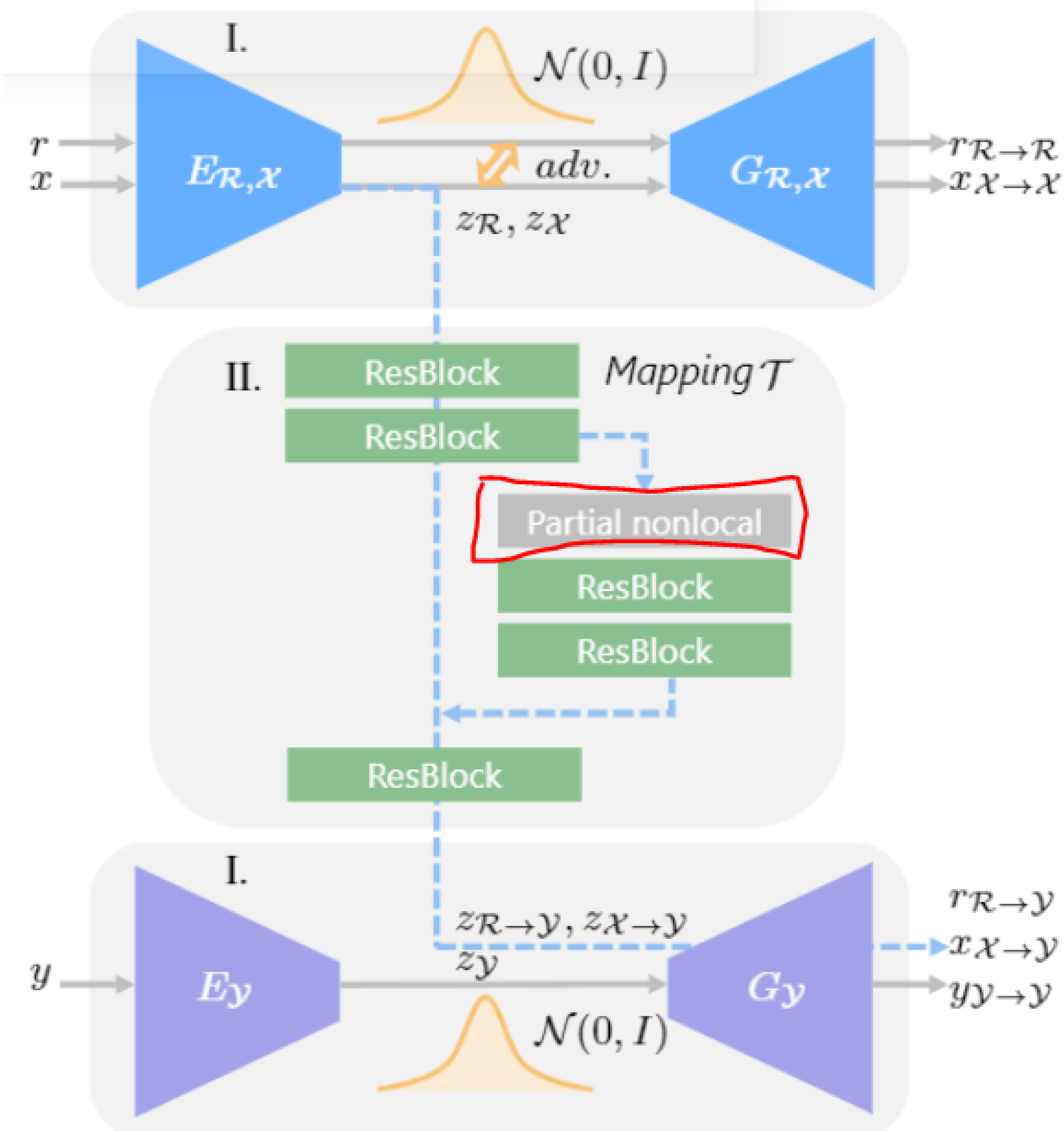
- specifically matches the multi-level activations of the adversarial network D_M, and that of the pretrained VGG network (also known as perceptual loss) where$\phi^i_{D_T}$ (ϕ^i_{VGG}) denotes the i^{th} layer feature map of the discriminator (VGG network), and $n^i_{D_T}$ (n^i_{VGG}) indicates the number of activations in that layer.

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\sum_i \frac{1}{n^i_{D_\mathcal{T}}}\|\phi^i_{D_\mathcal{T}}(x_{\mathcal{X}\to\mathcal{Y}}) - \phi^i_{D_\mathcal{T}}(y_{\mathcal{Y}\to\mathcal{Y}})\|_1 + \sum_i \frac{1}{n^i_{\mathrm{VGG}}}\|\phi^i_{\mathrm{VGG}}(x_{\mathcal{X}\to\mathcal{Y}}) - \phi^i_{\mathrm{VGG}}(y_{\mathcal{Y}\to\mathcal{Y}})\|_1\Big],$$

Method #2: Multiple degradation restoration

- background:
  - The latent restoration(method #1) using the residual blocks, as described earlier, only concentrates on local features due to the limited receptive field of each layer
  - the restoration of structured defects requires plausible inpainting, which has to consider long-range dependencies so as to ensure global structural consistency

- Concept:



- enhance the latent restoration network by incorporating a global branch, which composes of a nonlocal block that considers global context and several residual blocks in the following.
  - nonlocal block: explicitly utilizes the mask input so that the pixels in the corrupted region will not be adopted for completing those area
  - partial nonlocal block

- painting as partial nonlocal block. Formally, let F∈RC×HW be the intermediate feature map in M(C, H and W are number of channels, height and width respectively), and m∈ {0,1}^{HW} represents the binary mask downscaled to the same size, where 1 represents the defect regions to be inpainted and 0 represents the intact regions. The affinity(s) between i^{th} location and j^{th} location in F, denoted by $s_{i,j}\in R^{HW\times HW}$, is calculated by the correlation of F_i and F_j modulated by the mask (1−m_j)