1. Pose estimation
    a. Predict 6 DOF pose based on 2D input data
    b. Input data can be RGB images and possibly depth
    c. Pose can be extracted e.g. by projecting 2D points onto 3D model
    d. Needs labeled data for training, ideally real
2. Real training data
    a. Hard to acquire, not many datasets available & hard to do manually
    b. Can lead to overfitting due to dataset being to specialized (illumination, environment, texture..)
3. Synthetic training data
    a. Easy to produce large amount of training data
    b. Easily leads to overfitting due to lack of noise (e.g. compression), visual differences between 3D model / real counterpart, crude illumination & shading
    c. Bridging the domain gap between real / synthetic therefore main challenge
4. Domain randomization (DeceptionNet)
    a. Randomize factors that the algorithm should not be sensitive to
    b. Training in two alternating phases:
        i. Freeze recognition, update deception weights to maximize the loss of the recognition NW
        ii. Freeze deception, update recognition weights to minimize the loss of the recognition NW
    c. This encourages deception NW to maximally confuse the recognition NW, and the recognition NW becomes increasingly more resistant to randomness
    d. Deception & augmentation of input data:
        i. Deception NW follows encoder / decoder method, feeding encoded vector to decoding modules
            1. Noise: Randomness to encoded vector
            2. Distortion: Elastically image deformations
            3. Light: Phong lighting (ambient, diffuse, specular) + Light direction
            4. Background: Upsampling & convolutions
    e. Overall, rather good at generalization because of independence from target domain
5. Photorealistic image synthesis
    a. Fully synthetic data generation from 15 objects / 6 scenes
    b. PhysX + Arnold to achieve high degree of realism
    c. Paper notes that PBR quality is instrumental for good training
6. Our approach
    a. Overview
        i. PhysX + Blender / AS for rendering objects
        ii. Scene is not rendered, real images used to stitch 3D models of scenes are used instead
        iii. Renderings + Images are blended intelligently
        iv. Composed image improved by means of harmonization (later)
    b. Discussion
        i. Minimizes domain gap by using real images as background
        ii. Lower costs & faster rendering due to not rendering the scene
        iii. Illumination needs to be accurately replicated
        iv. Camera pose + parameters must match original capture
        v. Scene geometry / physical presence needs to be reflected on objects without rendering it (e.g. shadows, indirect light)

c. Harmonization
    i. Input is RGB image + foreground mask
    ii. Encoder / decoder method with skip links to prevent loss of detail and the loss function preferring blurry images
    iii. Two decoders, one to reconstruct the harmonized output, one to parse the scene & predict semantic labels
    iv. Labels are used by the reconstruction decoder
    v. Encoder is shared by both decoders, both decoders have skip links
    vi. Semantic information helps color distribution of e.g. skin / sky and which regions to match for better adjustment
d. First results