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 annotated 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..)
    c. Example: LineMOD trained network struggles with easy inputs from HomebrewDB, probably overfitted
    d. Therefore, hard to generalize
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. Deception NW tries to maximally confuse the recognition NW,
Recognition NW becomes increasingly more resistant to randomness
    c. Training in two alternating phases:
        i. Freeze recognition, update deception, maximize loss of recognition NW
        ii. Freeze deception, update recognition, minimize loss of recognition NW
    d. Deception & augmentation of input data:
        i. Deception NW uses encoder / decoder method,
feeds encoded vector to decoding modules, which are
            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 / Appleseed for rendering objects
        ii. Scene is not rendered, real images used instead
        iii. Images are from the ones used to generate 3D models of scenes
        iv. Renderings + Images are blended intelligently
        v. Composed image improved by using harmonization (later)
    b. Discussion
        i. Advantages
            1. Minimizes domain gap by using real images as background
            2. Lower costs & faster rendering due to not rendering the scene
        ii. Challenges
            1. Illumination needs to be accurately replicated
            2. Camera pose + parameters must match original capture
            3. Scene geometry / physical presence needs to be reflected on objects without rendering it (e.g. shadows, indirect light)
    c. Possible solutions
        i. Extracting light information from image (DeepLight)
        ii. Harmonization, improving the image composition
    d. Harmonization
        i. Input is RGB image + foreground mask
        ii. Encoder / decoder method with skip links to prevent loss of detail & the loss function should not prefer blurry images
        iii. Two decoders, one to reconstruct the harmonized output, one to parse the scene & predict semantic labels
        iv. Encoder is shared by both decoders, both decoders have skip links
        v. Labels are used by the reconstruction decoder
        vi. Semantic information helps color distribution of e.g. skin / sky & which regions to match for better adjustment
    e. First results