Style Swap

Brain & Cognitive Society, IIT Kanpur

Summer Project 2024

Abstract

FICE enhances fashion image editing using StyleGAN2, CLIP, e4e encoder, segmentation, and DensePose. It allows realistic clothing integration based on text descriptions, avoiding the need for target images. Through rigorous experiments, FICE achieves superior performance in generating convincing fashion images while preserving identities, surpassing previous methods.

Segmentation Model

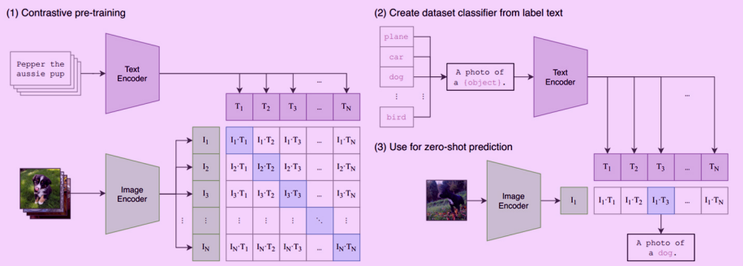
It uses DeepLabv3 to segment images into background, head, and body masks. It computes two semantic losses: *Image loss* preserves background and face regions, crucial for skin tone and color fidelity. *Head Loss* ensures hair preservation and addresses interactions between head and body regions not covered by Image loss. These losses enhance the realism and fidelity of generated fashion images.

Body Segmentation Mask

Head Segmentation Mask

CLIP Model

CLIP (Contrastive Language-Image Pretraining) by OpenAI is a neural network that learns to associate text descriptions with images. It uses a Transformer-based text encoder and a CNN-based image encoder to produce embeddings of visual and textual data.

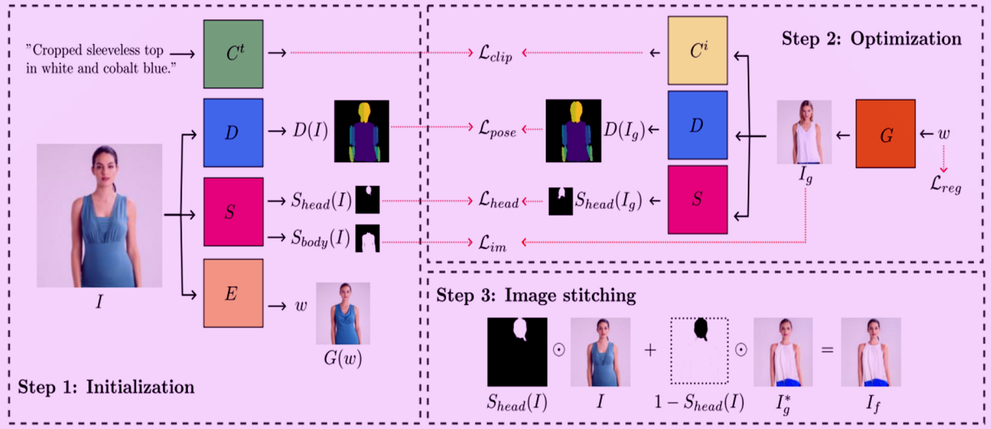


Trained on 400 million image-text pairs, CLIP employs contrastive learning to maximize similarity between correct pairs and minimize it for incorrect ones. This enables zero-shot learning, allowing the model to perform tasks without task-specific training.

Code:-
https://github.com/qu-bit1/style-swap
Documentation:-
https://shorturl.at/ejtwF

Methodology

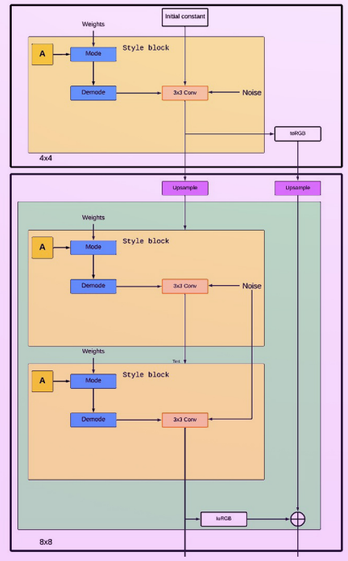
The FICE project's main loop generates fashion images from text by encoding text and images into e4e's latent vectors. CLIP embeddings guide modifications before StyleGAN2 creates initial images. Segmentation and pose preservation ensure semantic accuracy. Losses refine latent vectors via backpropagation over epochs, enhancing image quality and alignment with descriptions.



"Long Sleeve Yellow Shirt" ***"Sleeveless T-Shirt in Purple"***

StyleGAN-2

StyleGAN2 by Nvidia is a high-level GAN for superior image synthesis, refining its precursor with better feature control. Its architecture integrates a mapping network, generator, and discriminator to produce detailed, realistic images.

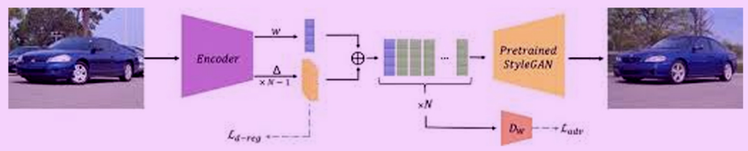


The Generator:
The StyleGAN2 generator starts with a constant and progressively doubles the feature map resolution through a series of blocks. Style blocks use the latent space w to adjust and normalize weights, enhancing image quality and variety. The final image is formed by scaling and summing the RGB outputs of all blocks, added random noise for detail.

The discriminator:
It is a CNN that evaluates images at multiple scales to distinguish real from generated ones. It utilizes residual connections and progressive growing. This architecture enables it to extract hierarchical features and provide precise feedback to the generator.

e4e Encoder

It converts input images into a disentangled latent space w of 512 dimensions from an initial latent vector z.



Ihis w vector is stacked to form a w+ vector of size (#_synthesis_blocks_in_StyleGAN2, 512). During inference, w+ vectors are adjusted by backpropogation, which optimizes for manipulation and generates images aligned with specified editing parameters.

Experimentation Code:-
https://www.github.com/IdhantKadela/Style-Swap
https://www.github.com/Vedant336Neekhra/gfe
https://github.com/rosinality/stylegan2-pytorch