

HYBRID SYNTHETIC PERSON DATASET GENERATION FOR CCTV SURVEILLANCE USING DIFFUSION MODELS AND INTELLIGENT COMPOSITION

Sude Naz Öztürk, 2220765041

Barış Çelik, 2220765033



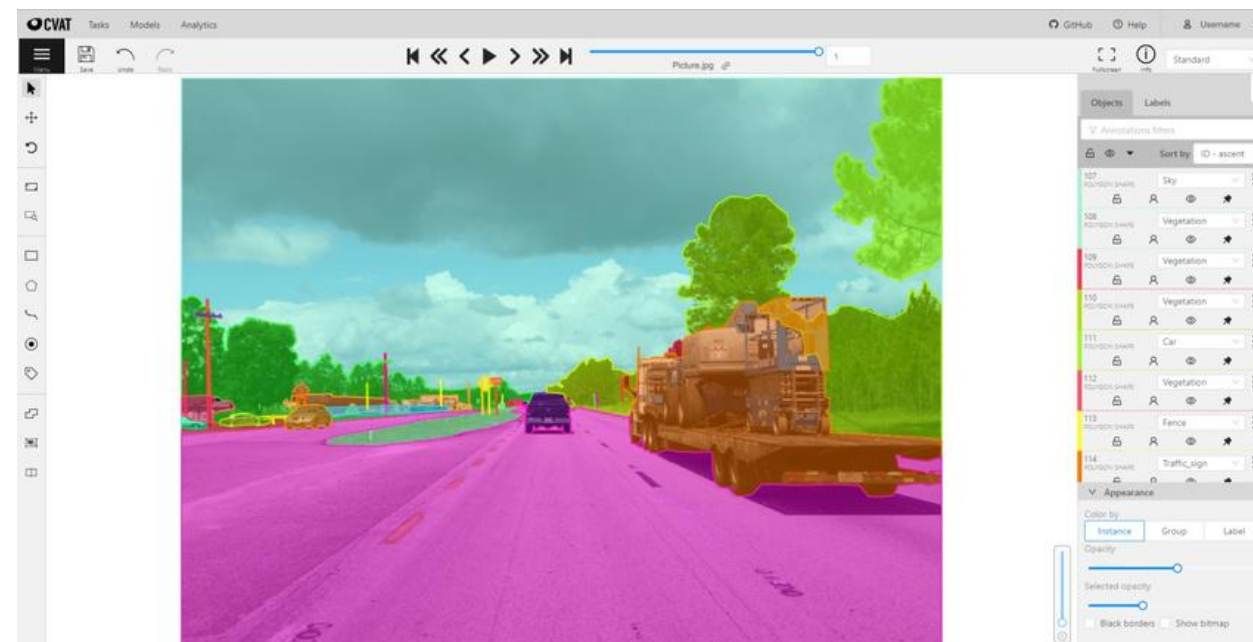
Agenda

| | |
|----------|-----------------------------------|
| <i>1</i> | Introduction |
| <i>2</i> | Related Work |
| <i>3</i> | Dataset |
| <i>4</i> | Model & Methods |
| <i>5</i> | Training Details |
| <i>6</i> | Experiments |
| <i>7</i> | Results |
| <i>8</i> | Limitations & Critical Evaluation |
| <i>9</i> | Future Work |

Introduction

THE PROBLEM:

- Person detection models require **large labeled datasets**
- Manual annotation is **expensive** and **time-consuming** 😞 😞
- **Privacy concerns** with real surveillance footage
- **Limited diversity** in existing datasets



Our main goal is
to
get rid of labeling.
It's boring

Introduction

Our Solution

- Generate synthetic persons using Diffusion Models
- Intelligently compose them into real CCTV backgrounds
- Create hybrid datasets (real + synthetic persons)

Why This is Matter

- **Real- World Impact:** Smart Cities, Retail Analytics, Security Systems, Autonomous Vehicles.



Introduction

Project Novelty

- Custom Diffusion Model
- Intelligent Composition Pipeline
- Two class hybrid approach
- Domain Specific Optimization



Generated
persons

+



Background scene +
masks



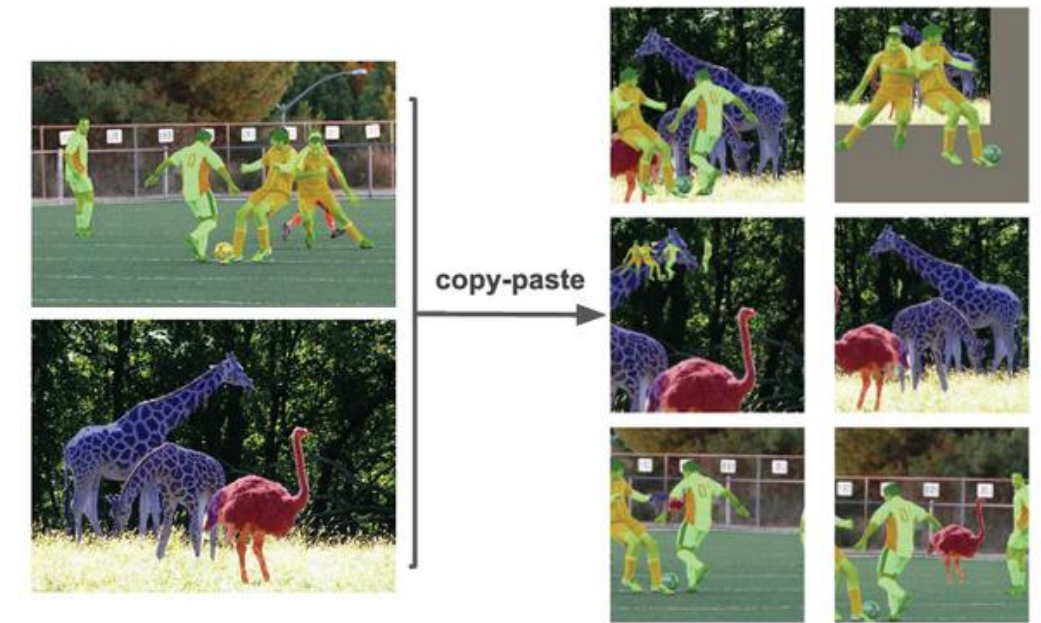
After full pipeline

RELATED WORKS

1) Copy-Paste Augmentation (Google, CVPR 2021)

Cut real persons → Paste randomly

- Simple, effective (+2-3% mAP)
- Random placement, no perspective awareness, limited to real person cutouts



2) Cut, Paste and Learn (Facebook, ICCV 2017)

Object composition with blending

- Achieves ~95% of real data performance
- Random placement, no domain specificity

3) Pedestrian Synthesis(CVPR 2018)

GAN-based person synthesis for driving

- Domain-specific (autonomous driving)
- Custom GAN training required, Limited to driving scenarios



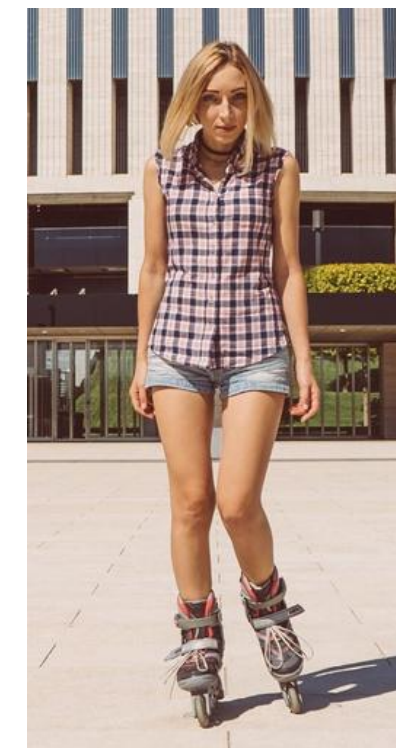
DATASET: FIRST-STAGE

1. Diffusion Model Training Data

- SHHQ-1.0 (Stylish Humans High Quality)
- Public research dataset
- RGB images + **segmentation masks**
- ~40,000 high-resolution person images(512x1024), and 40,000 masks

Characteristics

- **Purpose:** Train custom diffusion model
- Full-body persons
- Diverse poses, clothing, demographics
- Mostly Female(Not good for diversity)



(samples from SHHQ)

DATASET: SECOND STAGE

1. Composition Data

- **CCTV Background Frames**
- **Source:** [Youtube](#) videos and Google Images
- **Type:** **RGB** images (JPEG) **Size:** **100 frames**
- **Original Resolution:** **1280×720** (HD)
- **Final Resolution:** 640×480 (VGA)

Characteristics

- **Real CCTV** perspectives
- **Varied** scenes, lighting etc.
- Different camera angles
- Some **contains real persons**

PART 1 - Model & Methods

Why Diffusion Models?

- Stable training without mode collapse
- High sample diversity
- Strong preservation of global structure (human silhouette & pose)
- Well-suited for complex data distributions such as full-body humans

Why not GANs or VAEs?

- GANs: sharp images but unstable training and mode collapse risk
- VAEs: stable but overly smooth and blurry outputs

PART 1 - Model & Methods

Architecture

- Mask-conditioned DDPM with a U-Net backbone
- Input: noisy RGB image + binary human segmentation mask
- Output: predicted Gaussian noise at each timestep

Key Components

- U-Net with encoder–decoder and skip connections (standart for diffusion models)
- Residual blocks for stable deep training
- Group Normalization (robust for small batch sizes)
- SiLU activation for smoother gradients (standart for diffusion models)
- Sinusoidal time embeddings injected into all ResBlocks

PART 1 - Model & Methods

Conditioning Strategy

- Binary segmentation mask concatenated as an extra channel
- Improves silhouette consistency and reduces background artifacts

Loss Function

- Combination of MSE and L1 loss for noise prediction
- Face-weighted loss to enhance upper-body and facial details

PART 1 - Training Details

Optimization Setup

- Optimizer: Adam
- Learning rate: $2e-4$ (for stable convergence)

Training Configuration

- Epochs: 50
- Batch size: 32 (selected based on GPU memory limits)

Hardware

- GPU: NVIDIA A100
- GPU Memory: 80 GB
- System RAM: ~167 GB
- Disk: ~235 GB

PART 1 - Experiments



Experiment 1 — Baseline DDPM (From-scratch)

- **Setup:** Basic DDPM sampling, linear beta schedule
- **Observation:** Unstable visual quality (overexposed / faded samples), incomplete denoising artifacts

PART 1 - Experiments



Experiment 2 — EMA for Stabilization

- **Change:** Added Exponential Moving Average (EMA) weights for sampling
- **Why:** Reduce training oscillations and produce more consistent generations
- **Result:** More stable silhouettes and less fluctuation across samples

PART 1 - Experiments



Experiment 3 — DDIM Sampling (Inference-time improvement)

- **Change:** Replaced DDPM sampling with DDIM for faster, more deterministic sampling
- **Result:** Significant reduction in “white explosion / overexposure” issue

PART 1 - Experiments



Experiment 4 — Noise Schedule & Sampling Tweaks

- **Changes:**
 - Reduced `beta_end`
 - Increased EMA decay
 - Added `x0_pred` clamp in DDIM
- **Result:** Cleaner denoising and improved brightness stability

PART 1 - Experiments



Experiment 5 — Training Stabilization & Capacity

- **Changes:**
 - Gradient clipping
 - Added Self-Attention (capacity boost)
 - Switched loss to MSE + L1 (better learning signal)
- **Result:** Sharper body structure and fewer “melted” regions (but attention increased training cost)

PART 1 - Experiments



Experiment 6 — Incremental Training Refinements

- **Changes:**
 - Adjusted loss weighting between MSE and L1
 - Tuned EMA decay $0.9995 \rightarrow 0.9997$
- **Observation:** Incremental improvements in visual consistency, without a single dominant configuration

PART 1 - Experiments



Experiment 7 — Segmentation Mask Conditioning

- **Change:** Added binary human segmentation masks as an additional conditioning channel
- **Result:**
 - Significantly improved silhouette consistency
 - Fewer anatomical distortions (e.g., missing limbs, broken proportions)

PART 1 - Experiments

Final Experiment — Face-Weighted Loss & Extended Training

- **Change:** Introduced face-weighted loss to emphasize facial regions during training
- **Method:**
 - Higher loss weight applied to the upper body / head region
 - Combined MSE + L1 loss with spatial weighting
- **Training Update:**
 - Extended training to 50 epochs

PART 1 - Experiments



Epoch
5



Epoch
15



Epoch
25



Epoch



Epoch

PART 1 - Results

Final Results:



PART 2 - Dataset Generation Pipeline

Step 1: Generate Synthetic Persons

- Generate 500 persons using our diffusion model(192×512)
- Problem: Low resolution persons! Upscale the generated persons

Step 2: Collect and Resize Backgrounds

- 100 CCTV frames(downscale to 640×480)
- Output: predicted Gaussian noise at each timestep

Step 3: Annotate Walkable Areas

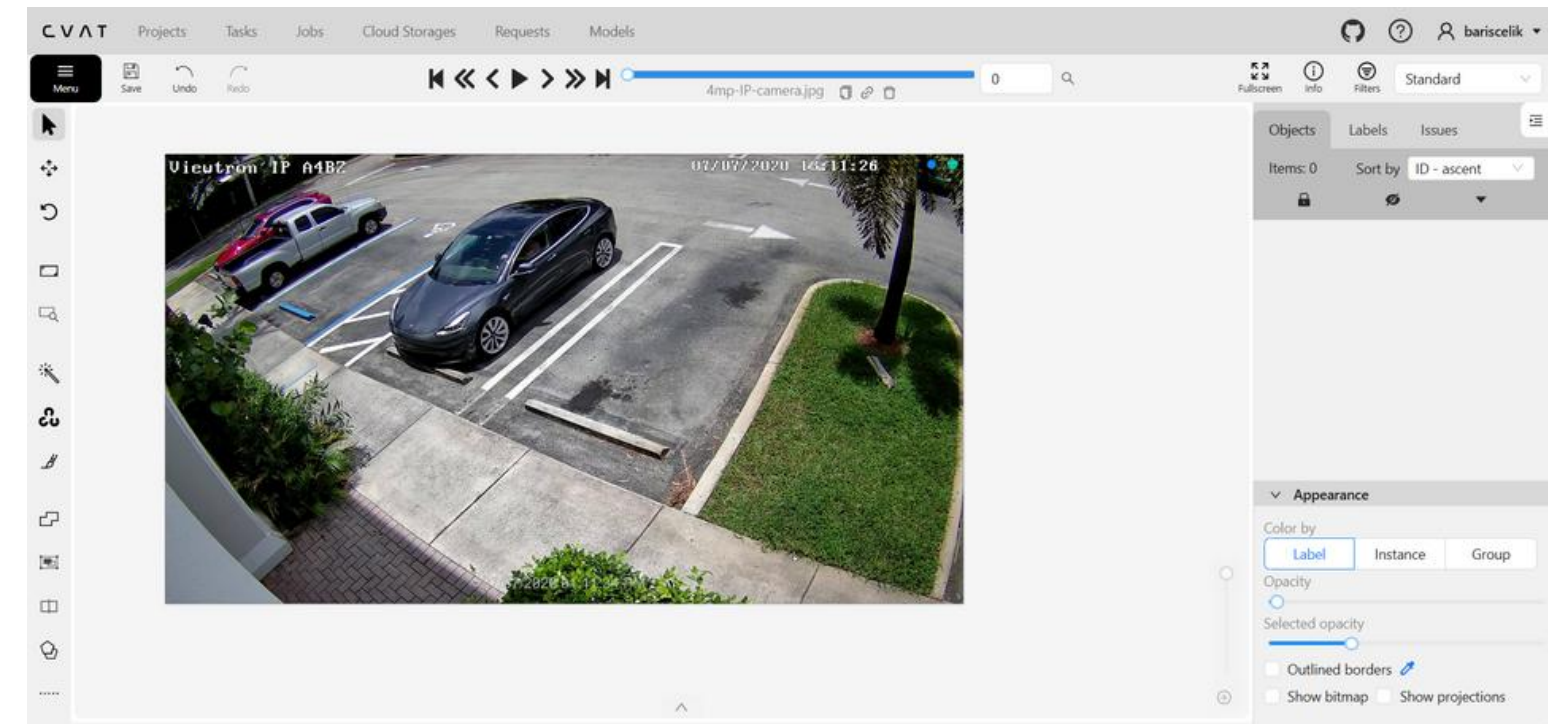
- We previously used pretrained SegFormer and SAM, but the results was not good!
- Used manual labeling on CVAT for perfect masks

Step 2-3

Pipeline



extract_frames.py
0.2 fps(every 5 second)



manually checking and masking
the walkable areas on CVAT(took
30 minutes for 100 images)

PART 2 - Dataset Generation Pipeline

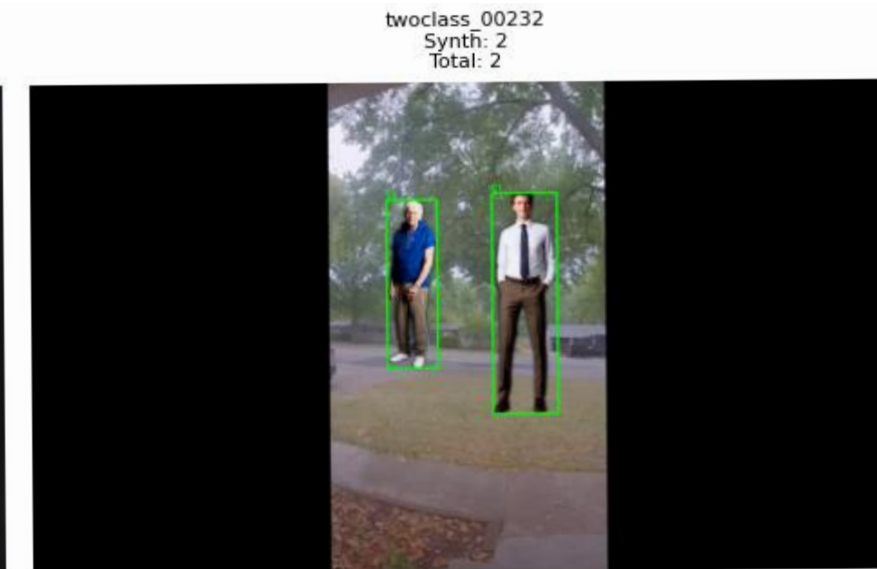
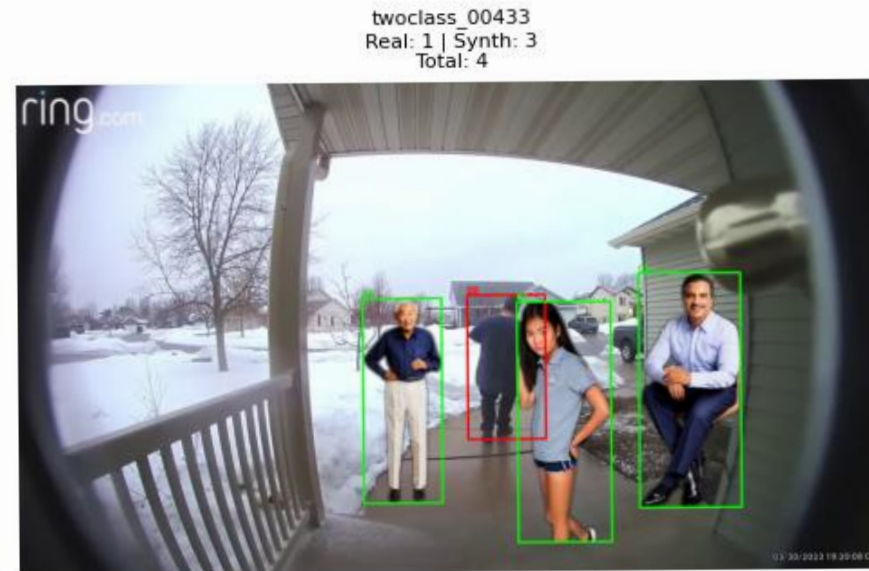
Step 4: Compose and Generate Dataset

- Place 1-3 synthetic persons per background
- Apply perspective scaling (0.5-0.8)
- IoU-based overlap prevention ($<5\%$)
- Feathered alpha blending(Softening the edges of the mask using Gaussian blur)
- Save labels in txt file(class x_center y_center width height)
- Run pretrained YOLOV11L (after composition) → detect real persons
- Merge labels: Real (Class 0) + Synthetic (Class 1)

RESULTS



Bounding Boxes



red: real persons
green: synthetic

Limitations & Critical Evaluation

- Collecting real background frame still need human
- Hard to find **cropped** diverse open-source **person dataset**
- Pretrained Models like SAM, SegFormer did not worked well, **human still need to mask walkable areas on background**
- Diffusion output is small (192x512 pixels), **Colab GPU Constraints**
- **Limited Pose Diversity** on generated persons
- **Slow Composition Pipeline: 500 image dataset take ~30-40 minutes**
- **Simplified Perspective Model:** Perspective is sometimes unrealistic in complex scenes

Strengths & Weaknesses

Strengths

- Cost effective
- Domain-Specific Optimization
- Intelligent Composition Pipeline
- Production-Ready Pipeline(End-to-end automated)

Weakness

- Diffusion Quality **Below SOTA**
- No Lighting Adaptation
- Static Composition
- **Manual Steps Required**
- **No Quantitative Validation yet**

A Failure



Pretrained Model Results



Future Work

- Quantitative Validation on generated dataset, calculate mAP
- Expand background dataset
- Still need to improve the **Diffusion Model**
- Automate Walkable areas

NEW APPROACHES

- Train a full diffusion model end-to-end for generating both persons and backgrounds together
- +++ **Synthetic Video Generation(COMPLEX)**

Thank You For Listening

*ANY
QUESTIONS*
?

