Few-Shot Adaptive Video-to-Video Translation

Ting-Chun Wang

NVIDIA

Recall the Motion Transfer Example

Behind the Scenes…

Disadvantages of vid2vid

- Separate models for each dataset

model 1

model 2

model 3

- Generalizing to new persons requires

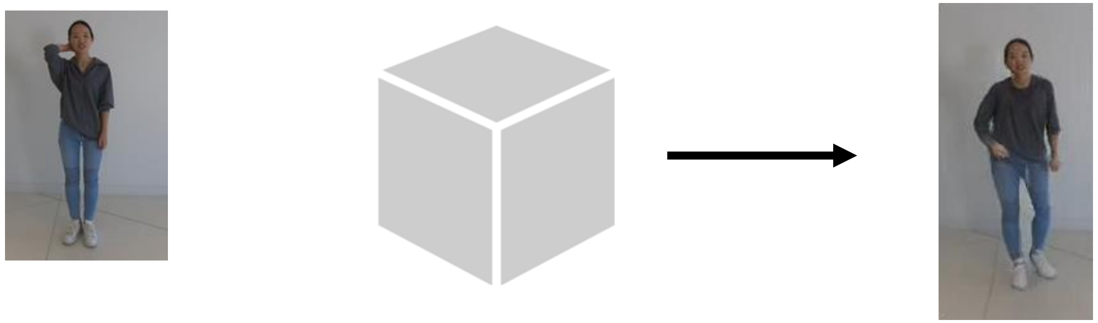
Collecting new data

Training

Wouldn’t it be great if…

- One model for all

- Dynamically determine the style at run time
  - based on an *exemplar image*

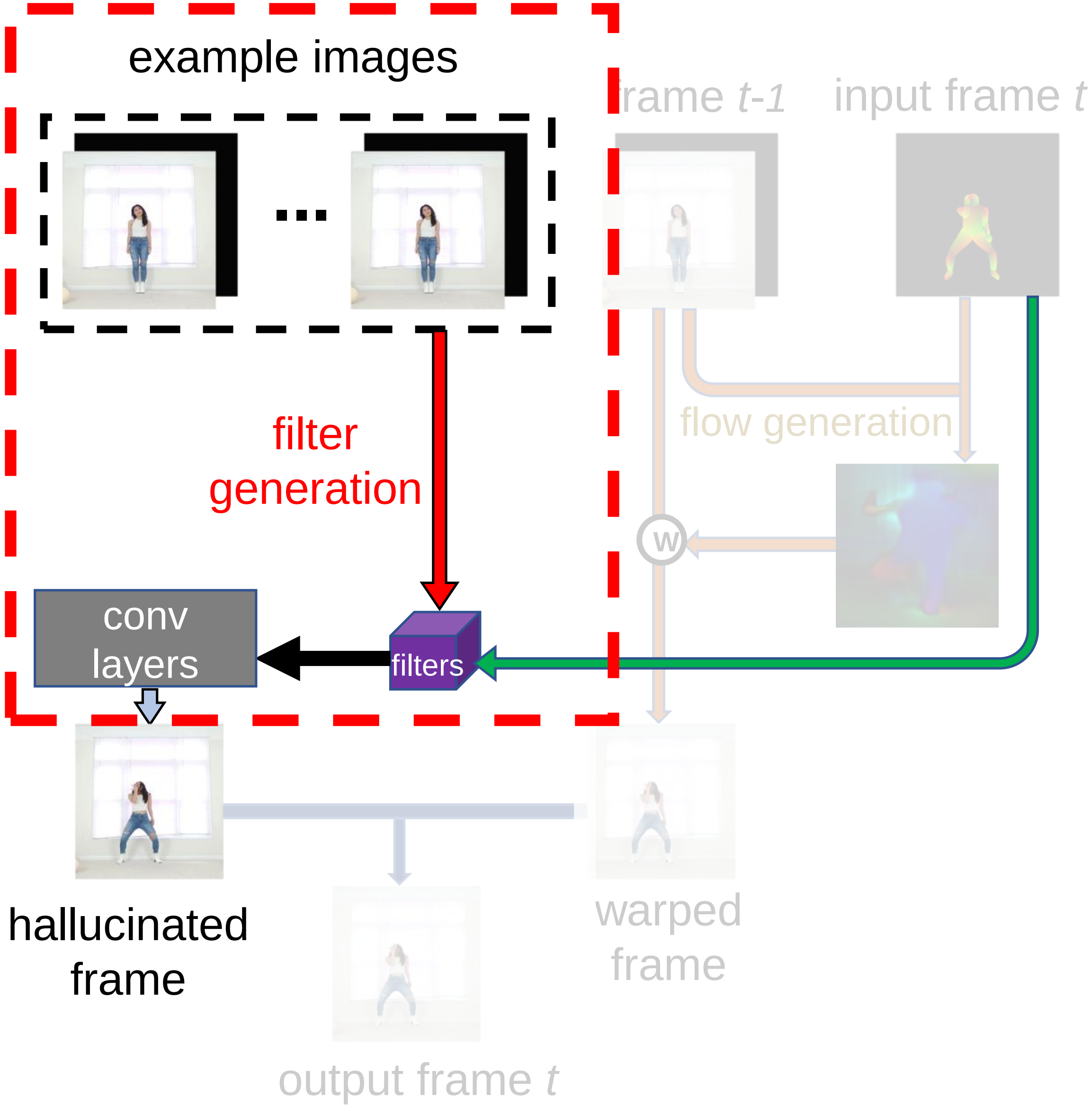


Adaptive Video-to-Video Translation

T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, B. Catanzaro, “Few-shot Adaptive Video-to-Video Synthesis,” To appear at NeurIPS 2019.

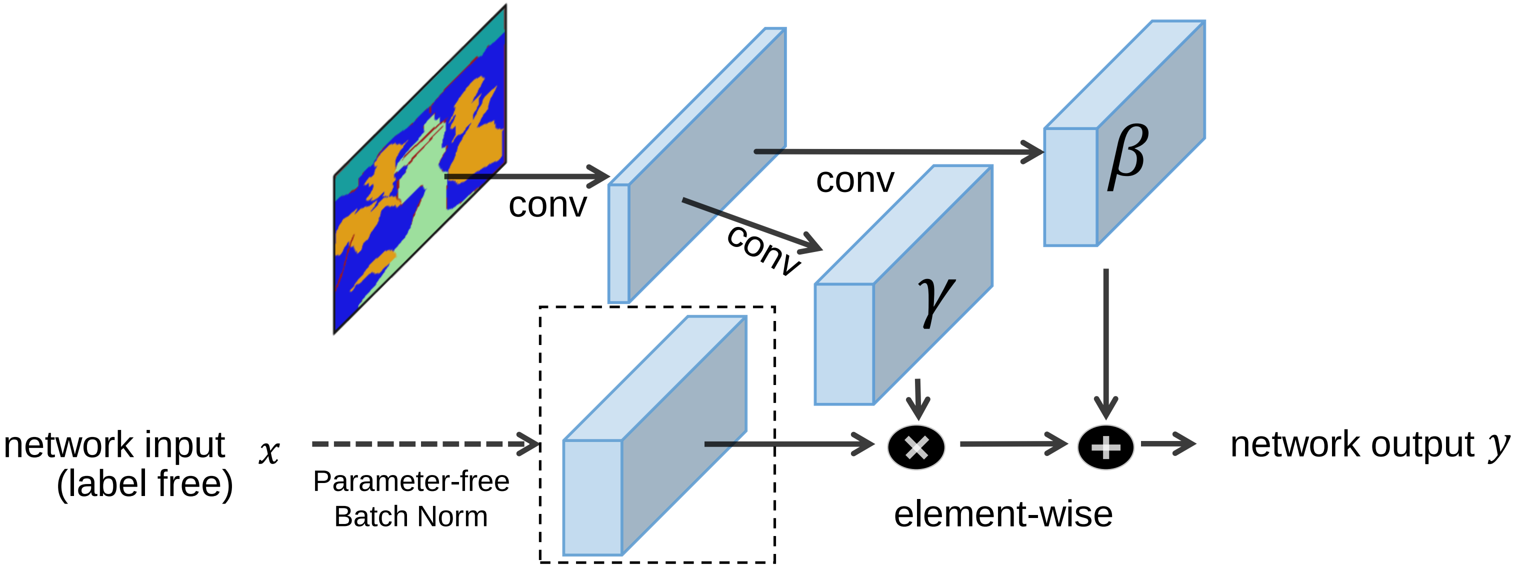
Adaptive vid2vid: overflow

- Original vid2vid
  - Output frame = Hallucinated frame + Warped frame
- Adaptive vid2vid
  - Hallucinated frames
    - generated based on example images
  - Using a filter generation scheme



Adaptive vid2vid

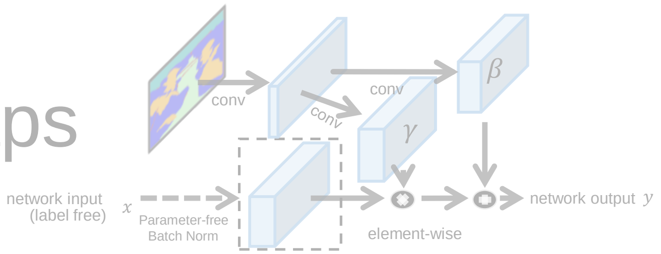
- Based on SPADE (GauGAN)
  - ~~Prior work: input semantics → encoder-decoder → output image~~
  - Instead: input semantics
    → ***spatially-varying*** normalization maps
    → used in every BatchNorm



$$y = \frac{x-\mu}{\sigma} \cdot \boxed{\gamma + \beta}$$

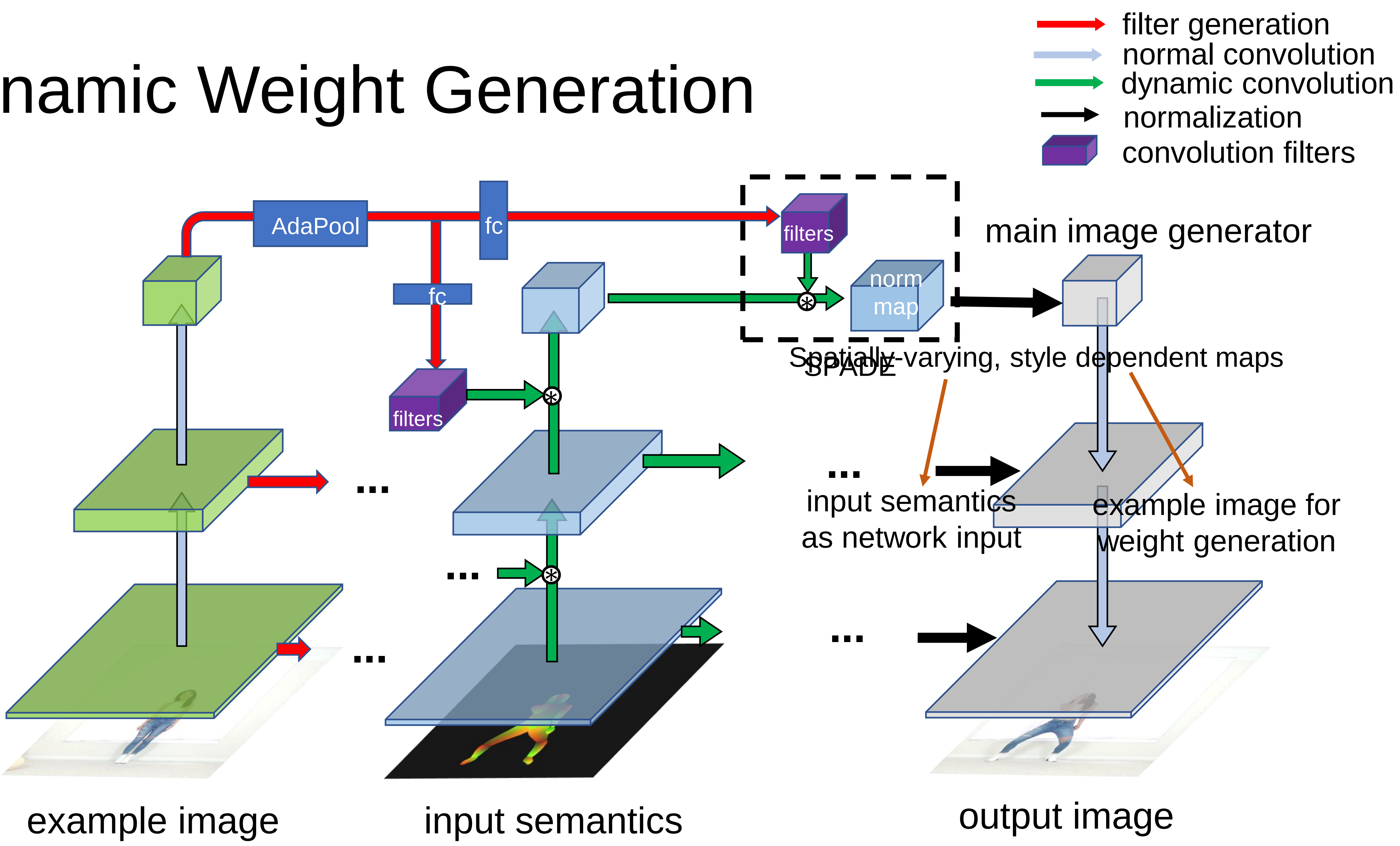
Adaptive vid2vid

- Based on SPADE (GauGAN)
  - Prior work: input semantics → encoder-decoder → output image
  - Instead: input semantics
    → ***spatially-varying*** normalization maps
    → used in every BatchNorm

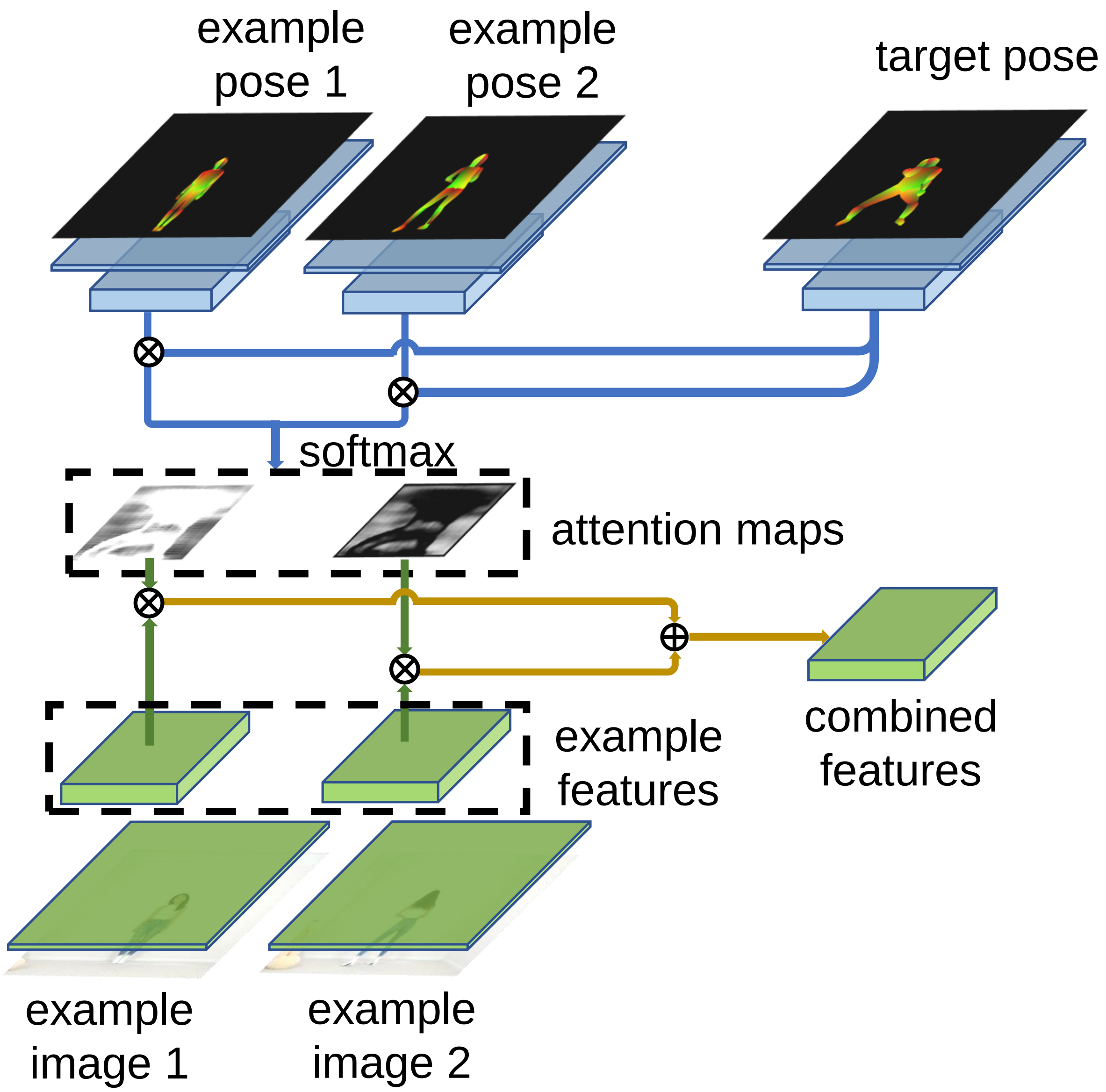


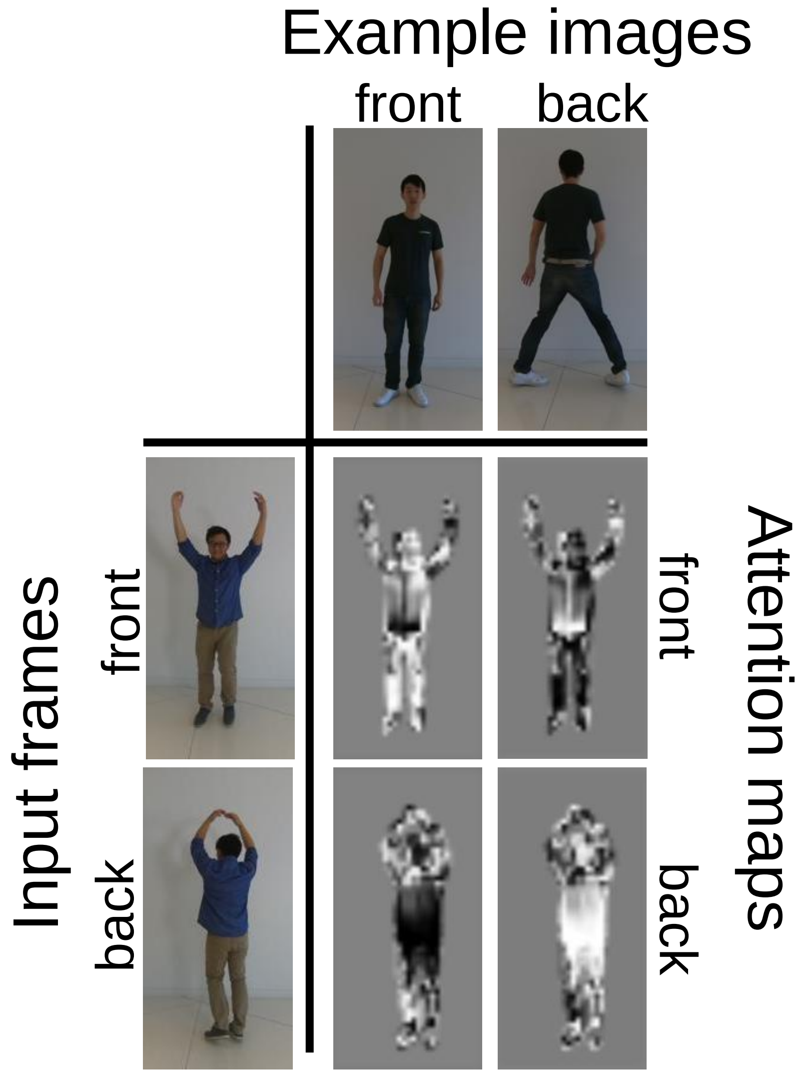
- Given an additional exemplar image
  - Dynamically configure the ***network weights*** in SPADE
  - Generate ***spatially-varying, style-dependent*** normalization maps
    - Spatial info ← input semantics
    - Style info ← exemplar images

Dynamic Weight Generation



Utilizing Multiple Example Images





Adaptive vid2vid: Training

- From a video
  - Randomly sample a clip
  - Randomly sample another reference frame(s)

- Make the network generate the clip
  - Based on the reference frame

Adaptive vid2vid: Testing

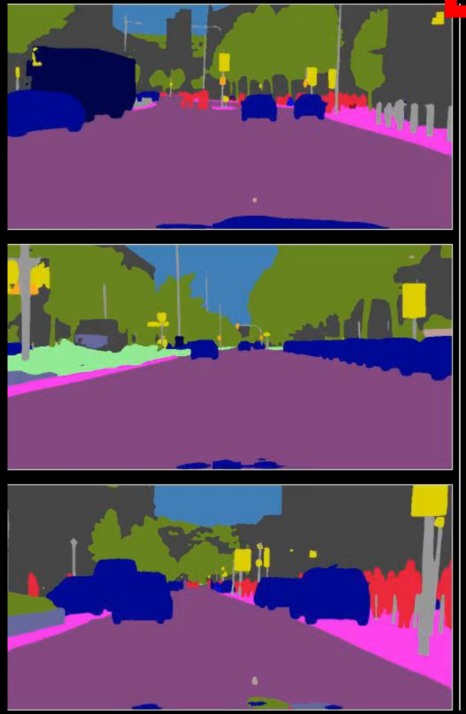
- Given an example image
- Finetune on the example image
  - Network output should be the same as the example
  - Only finetune for a few iterations

- For faces: normalize keypoints
  - To the same as example image
  - To better preserve identity

Results

- Semantic → Street view scenes
- Edges → Human faces
- Poses → Human bodies

Street View Scenes

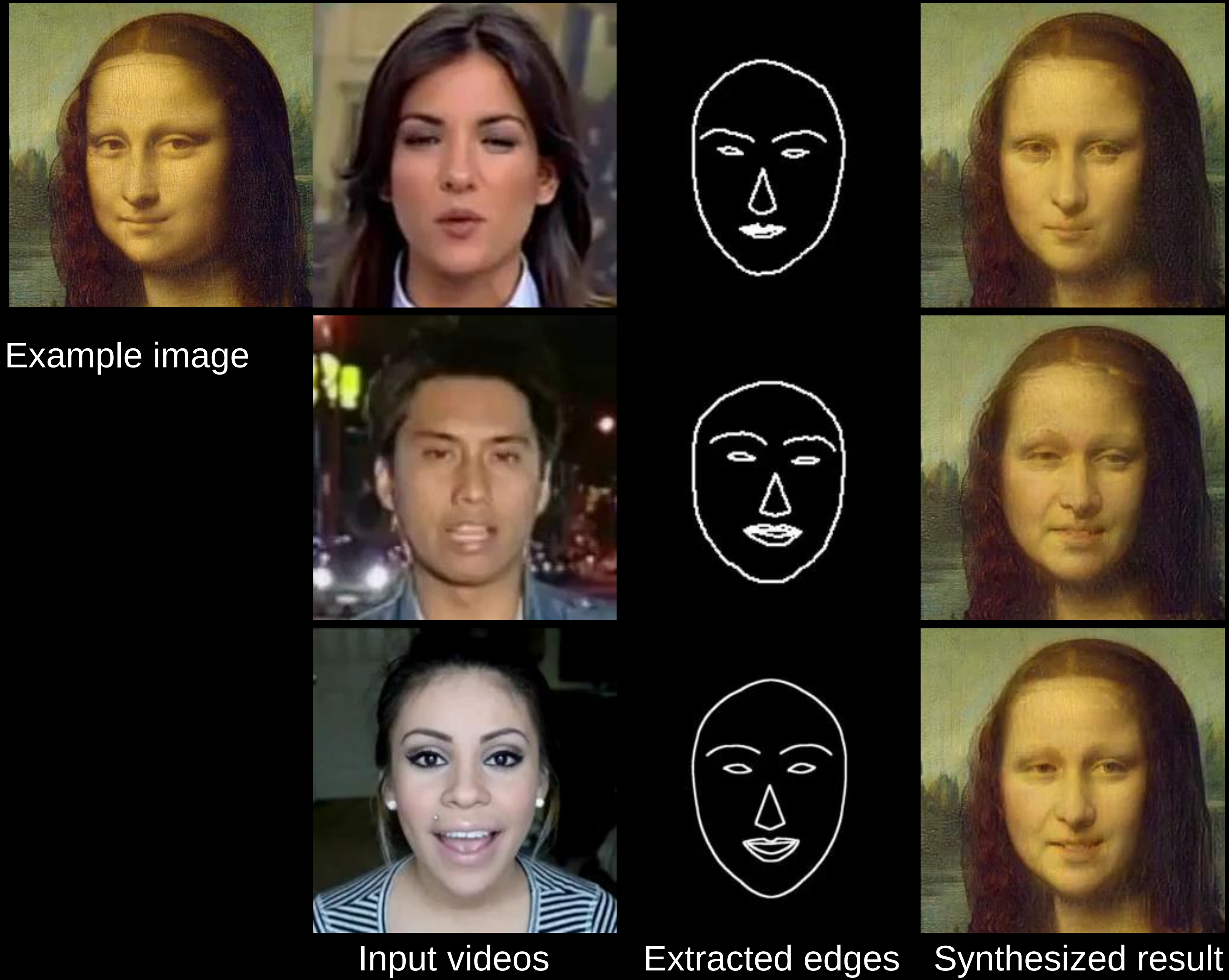


Input segmentations

Edges → Faces

Example images

Edges → Faces


Example image
Input videos
Extracted edges
Synthesized result

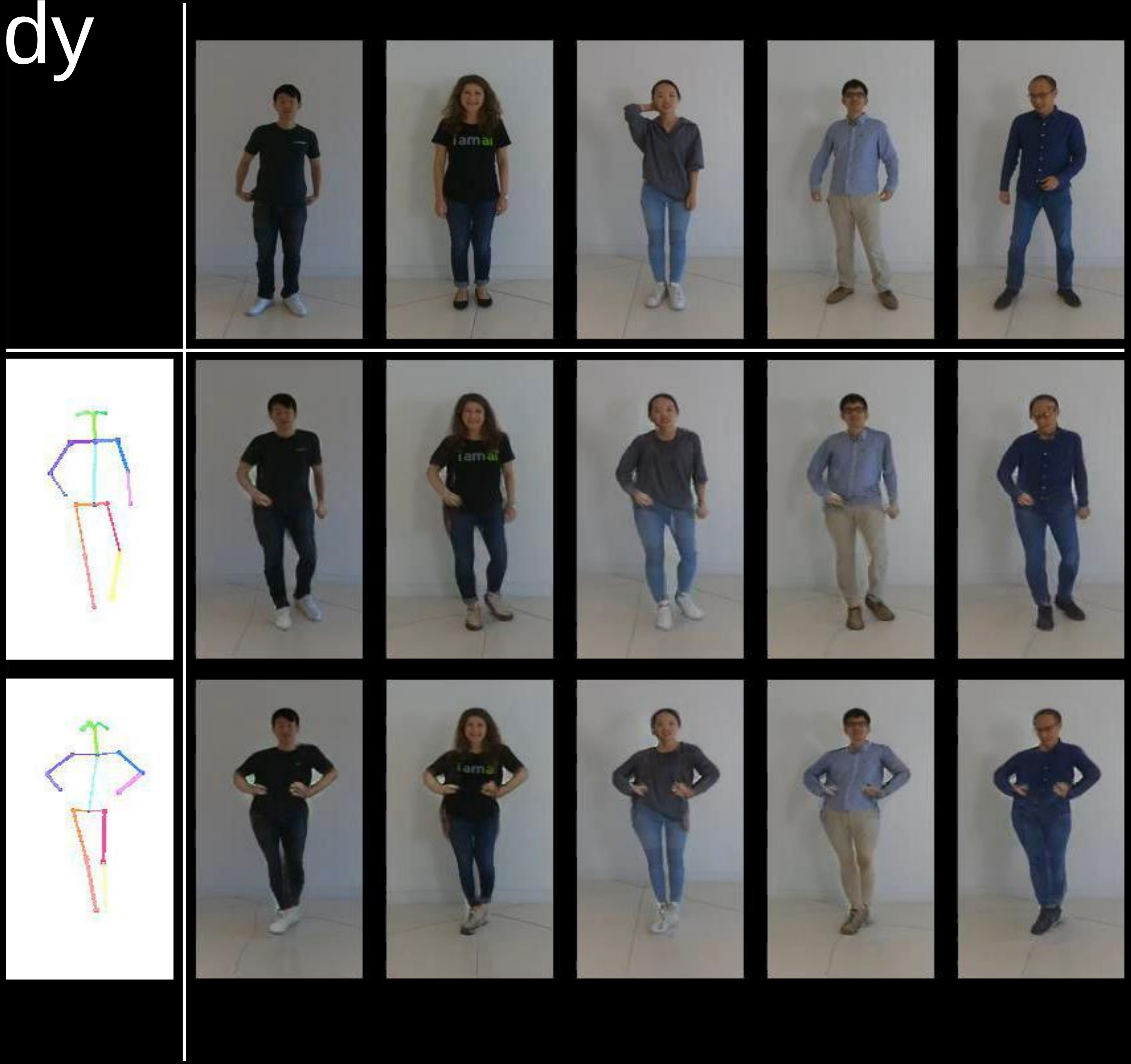
Poses → Body

Example images

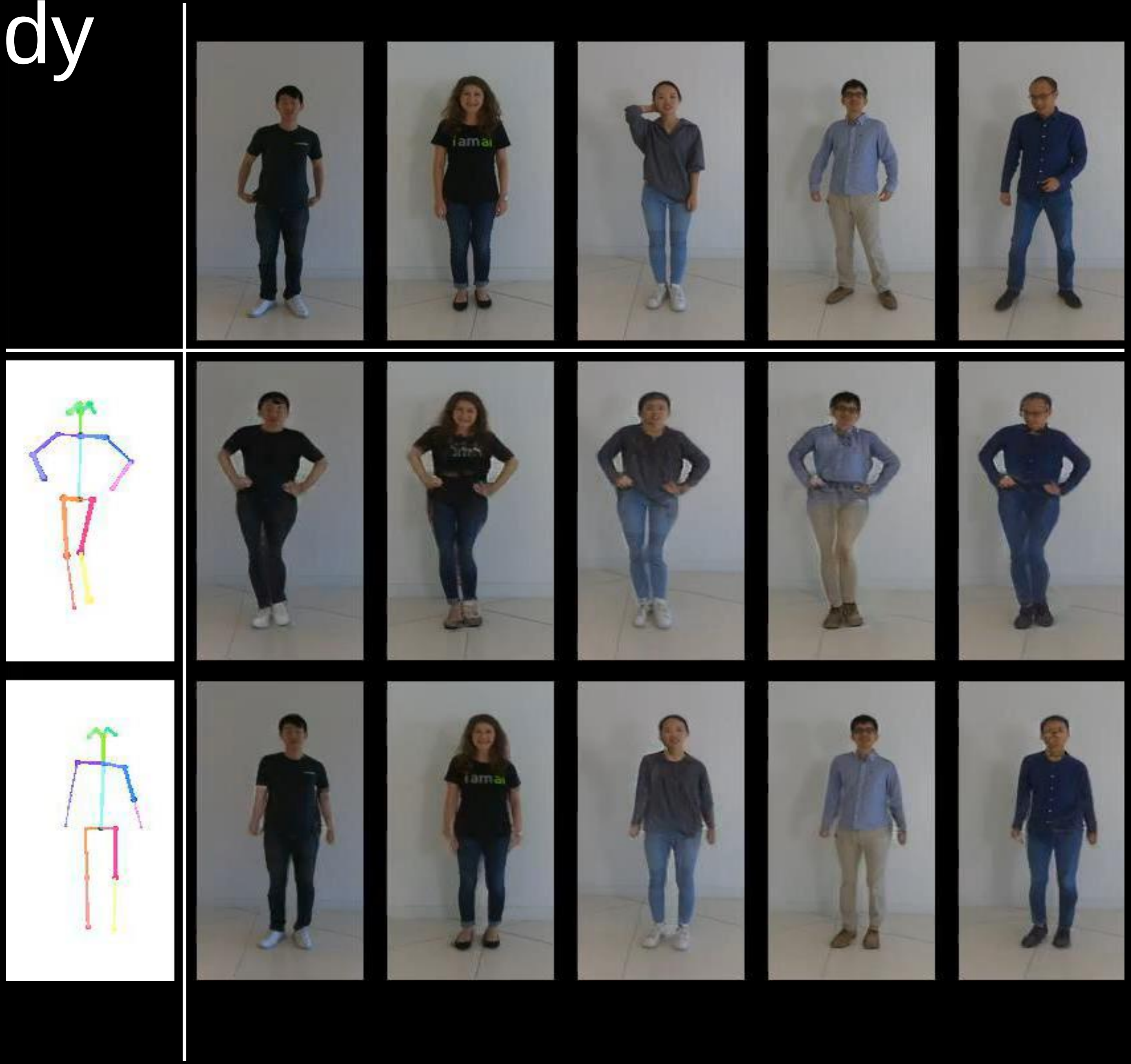
Input poses

Synthesized videos

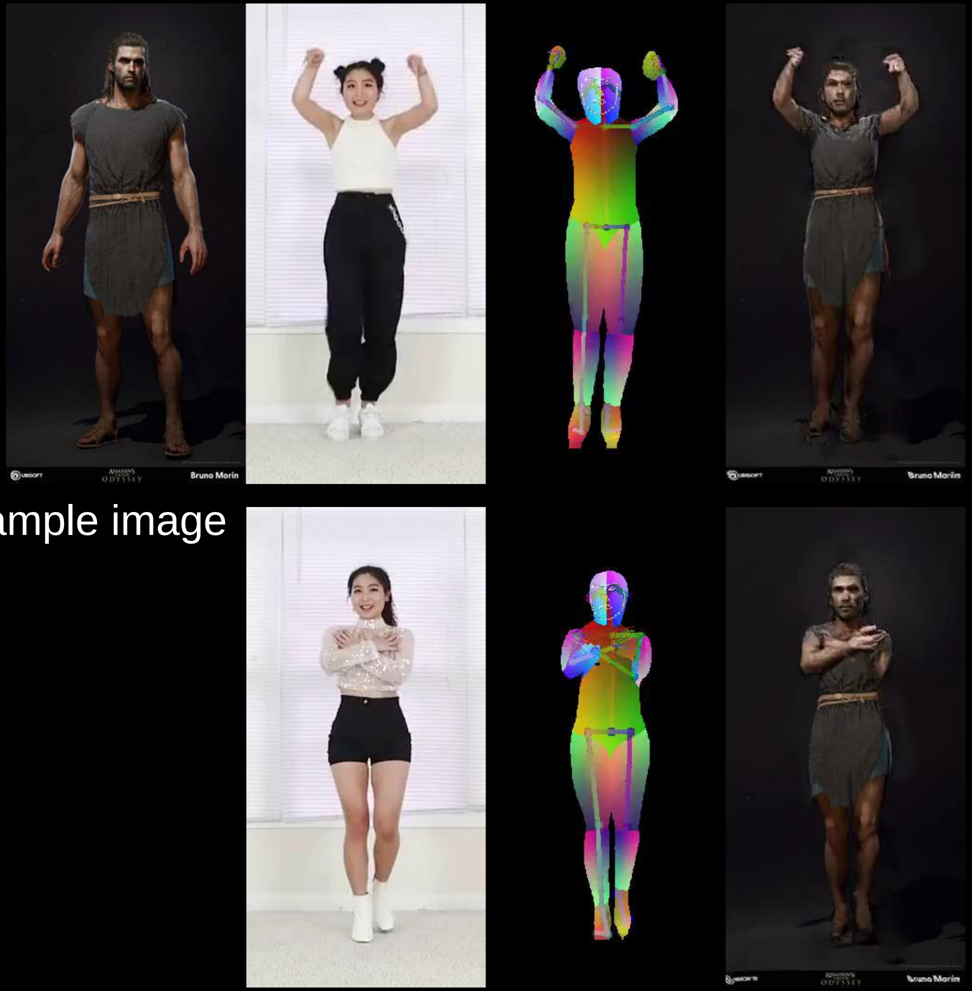
Poses → Body



Poses → Body



Poses → Body



Example image

Input videos | Poses | Synthesized

Conclusion

Conclusion

Generative adversarial networks (GANs)

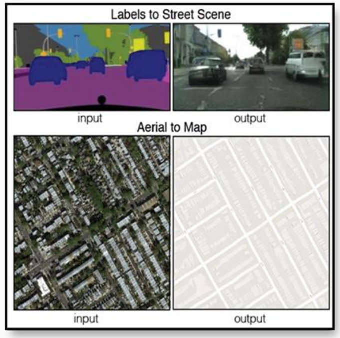
Unconditional $z \sim \mathcal{N} \rightarrow$ G $\rightarrow G(z) \rightarrow$ D $\rightarrow$

$x \in \mathcal{X} \rightarrow$ D $\rightarrow$

Conditional $x_1 \sim \mathcal{X}_1 \rightarrow$ F $\rightarrow F(x_1) \rightarrow$ D $\rightarrow$

$x_2 \sim \mathcal{X}_2 \rightarrow$ D $\rightarrow$

Supervised Image Translation



pix2pix *pix2pixHD* *GauGAN*

Unsupervised Image Translation

UNIT *MUNIT* *FUNIT*

Video Translation

vid2vid *vid2game* *Adap vid2vid*

THANK YOU

Questions?