Deformation Models for Image and Video Generation

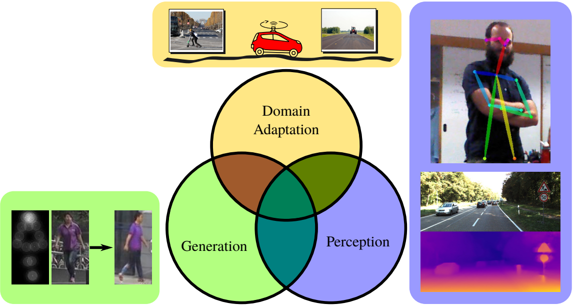
Stéphane Lathuilière

LTCI, Télécom Paris,
Institut polytechnique de Paris, France

Multimedia and Human Understanding Group (MHUG),
University of Trento, Italy

December 2020

Who am I?



PhD — Post-Doc — Assistant Professor

2014 — **Perception** Dr. R. Horaud

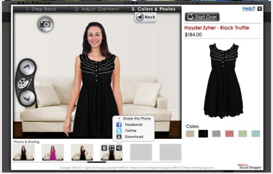
2018 — **MHUG** Prof. Nicu Sebe

2019 — **Multimedia** Prof. Marco Cagnazzo

Artistic / Editing / marketing purposes

- Photo editing

- Augmented reality [a]

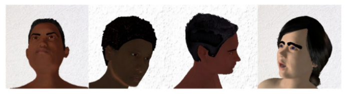


- Many other applications: video games, increasing some intrinsic image properties...

[a] image from *zugara* company

Machine Learning Tasks

- Generate annotated data: Head pose [1]

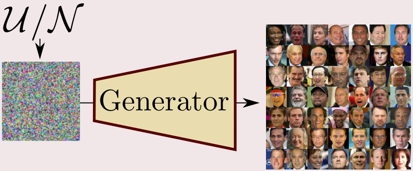


- Learning from few samples
- Domain adaptation

[1] **S.Lathuilère**, R.Juge, P.Mesejo, R.Munoz-Salinas, R.Horaud, Deep Mixture of Linear Inverse Regressions Applied to Head-Pose estimation, CVPR 2017

Image Generation

From Noise to Image



[4] P.Isola, J.-Y.Zhu, T.Zhou, A.A.Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR, 2017

Image Generation

From Noise to Image

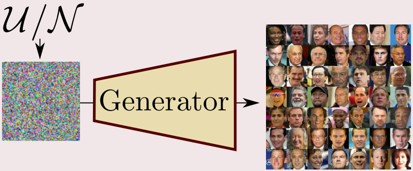
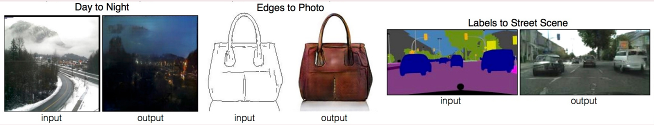
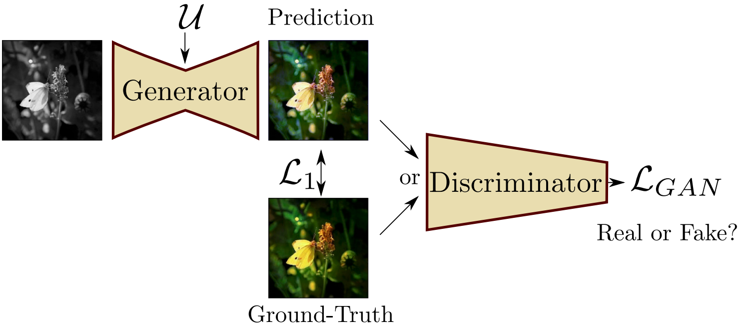


Image-to-Image translation [4]



[4] P.Isola, J.-Y.Zhu, T.Zhou, A.A.Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR, 2017



[4] P.Isola, J.-Y.Zhu, T.Zhou, A.A.Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR, 2017

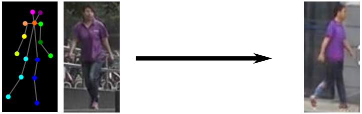
1. **Pose-based Human Image Generation**

2. **Multi-source Generation**

3. **Video generation: Image Animation**

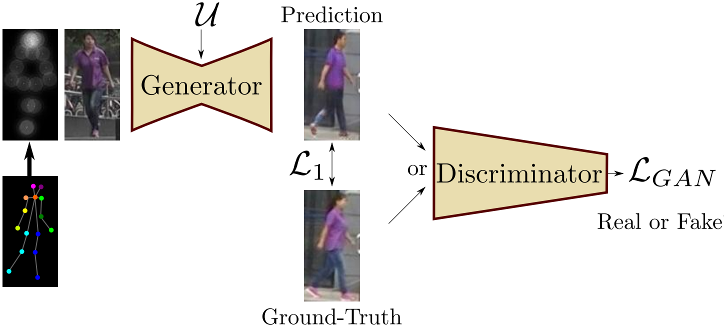
Pose-based Human Image Generation [5]

Prediction

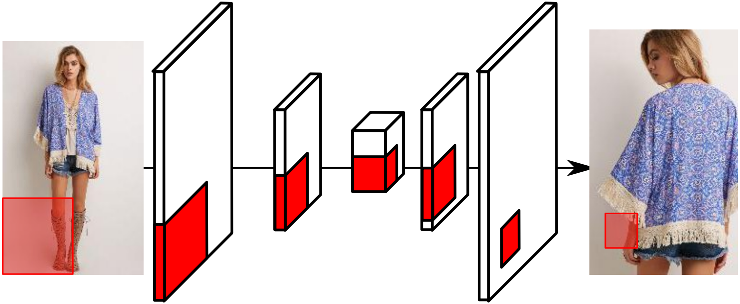


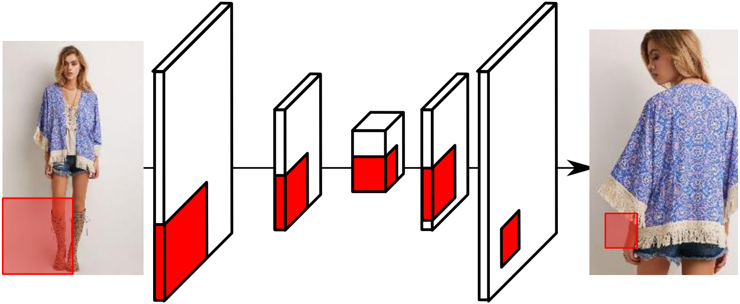
[5] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, Pose-guided person image generation, NIPS, 2017

Pose-based Human Image Generation [5]



[5] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, Pose-guided person image generation, NIPS, 2017

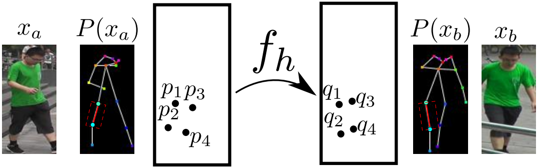




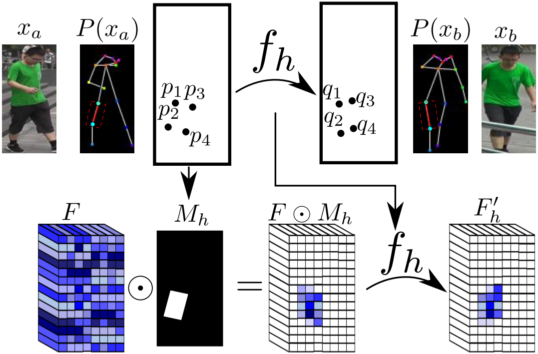
We need a deformation model!

Pose-based Human Image Generation: Our Proposal [6]

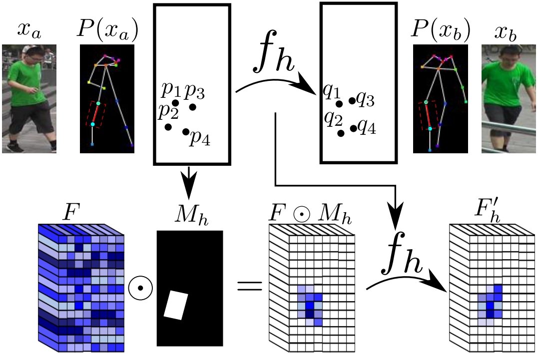
[6] A. Siarohin, E. Sangineto, **S. Lathuilière**, N. Sebe, Deformable GANs for Pose-based Human Image Generation, CVPR, 2018



[6] A. Siarohin, E. Sangineto, **S. Lathuilière**, N. Sebe, Deformable GANs for Pose-based Human Image Generation, CVPR, 2018



[6] A. Siarohin, E. Sangineto, **S. Lathuilière**, N. Sebe, Deformable GANs for Pose-based Human Image Generation, CVPR, 2018



The body parts are combined:

$$d(F) = max_{h=1,\ldots,10} F'_h, \quad (1)$$

[6] A. Siarohin, E. Sangineto, **S. Lathuilière**, N. Sebe, Deformable GANs for Pose-based Human Image Generation, CVPR, 2018

- $\mathcal{L}_1$ and $\mathcal{L}_2$ produce blurred images.

We propose a *nearest-neighbour* loss $\mathcal{L}_{NN}$

- Compute in a feature space $g(x)$.
- $g(\cdot)$: externally trained network.

$$\mathcal{L}_1^g(\hat{x}, x_b) = \sum_{\mathbf{p} \in g(\hat{x})} ||g(\hat{x})(p) - g(x_b)(p)||_1, \tag{2}$$

- $\mathcal{L}_1$ and $\mathcal{L}_2$ produce blurred images.

We propose a *nearest-neighbour* loss $\mathcal{L}_{NN}$

- Compute in a feature space $g(x)$.
- $g(\cdot)$: externally trained network.

$$\mathcal{L}_1^g(\hat{x}, x_b) = \sum_{\mathbf{p} \in g(\hat{x})} ||g(\hat{x})(p) - g(x_b)(p)||_1, \quad (2)$$

$$\mathcal{L}_{NN}(\hat{x}, x_b) = \sum_{\mathbf{p} \in g(\hat{x})} min_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} ||g(\hat{x})(\mathbf{p}) - g(x_b)(\mathbf{q})||_1, \quad (3)$$

- where $\mathcal{N}(\mathbf{p})$ is a $n \times n$ local neighbourhood of point $\mathbf{p}$

Pose-based Human Image Generation: ablation

x_a Target pose GT *Baseline* *Def.* *Full*

Figure: Qualitative results on the Market-1501 dataset.

Pose-based Human Image Generation: ablation

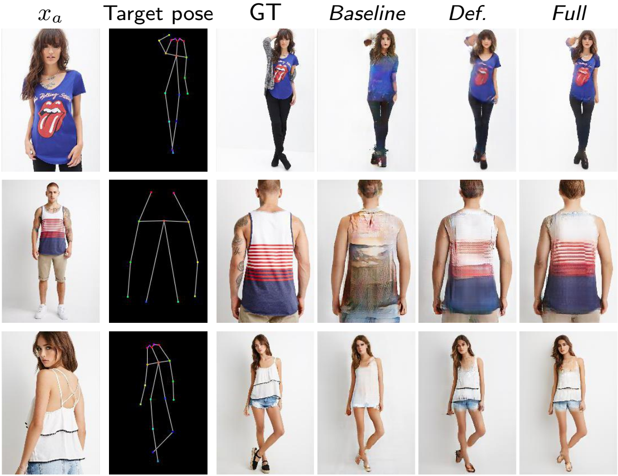


Figure: Qualitative results on the DeepFashion dataset.

Table: Comparison with the state of the art on the DeepFashion dataset.

Model	*SSIM*	*IS*
Ma et al. [7]	0.762	3.090
Ma et al. [8]	0.614	3.228
Esser et al. [9]	**0.786**	3.087
Ours	0.756	**3.439**

[7] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, Pose-guided person image generation, NIPS, 2017
[8] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, Disentangled person image generation, CVPR, 2018
[9] P. Esser, E. Sutter, and B. Ommer, A variational u-net for conditional appearance and shape generation, CVPR, 2018

	IDE + Euclidean [10] *Rank 1*	Discr. Embedding [11] *Rank 1*
No augmentation	73.9	78.3

Table: Data augmentation for Re-ID on the Market-1501 (*Rank 1* in %).

[7] L.Ma, X.Jia, Q.Sun, B.Schiele, T.Tuytelaars, and L.Van Gool, Pose-guided person image generation, NIPS, 2017

[9] P. Esser, E. Sutter, and B. Ommer, A variational u-net for conditional appearance and shape generation, CVPR, 2018

[10] L.Zheng, Y.Yang, and A.G.Hauptmann, Person re-identification: Past, present and future, arXiv, 2016

[11] Z. Zheng, L. Zheng, and Y. Yang, A discriminatively learned CNN embedding for person reidentification, TOMCCAP, 2018

[12] A. Siarohin, **S. Lathuilière**, E. Sangineto, N. Sebe, Appearance and Pose-Conditioned Human Image Generation using Deformable GANs, TPAMI, 2019

	IDE + Euclidean [10] *Rank 1*	Discr. Embedding [11] *Rank 1*
No augmentation	73.9	78.3
Ours (Full) [12]	**78.9**	**81.4**

Table: Data augmentation for Re-ID on the Market-1501 (*Rank 1* in %).

[7] L.Ma, X.Jia, Q.Sun, B.Schiele, T.Tuytelaars, and L.Van Gool, Pose-guided person image generation, NIPS, 2017

[9] P. Esser, E. Sutter, and B. Ommer, A variational u-net for conditional appearance and shape generation, CVPR, 2018

[10] L.Zheng, Y.Yang, and A.G.Hauptmann, Person re-identification: Past, present and future, arXiv, 2016

[11] Z. Zheng, L. Zheng, and Y. Yang, A discriminatively learned CNN embedding for person reidentification, TOMCCAP, 2018

[12] A. Siarohin, **S. Lathuilière**, E. Sangineto, N. Sebe, Appearance and Pose-Conditioned Human Image Generation using Deformable GANs, TPAMI, 2019

	IDE + Euclidean [10] *Rank 1*	Discr. Embedding [11] *Rank 1*
No augmentation	73.9	78.3
Ours (Full) [12]	**78.9**	**81.4**
Ours (Baseline)	68.1	70.6
Ma et al. [7]	66.9	73.9
Esser et al. [9]	58.1	63.1

Table: Data augmentation for Re-ID on the Market-1501 (*Rank 1* in %).

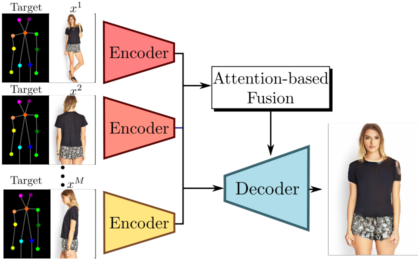
[7] L.Ma, X.Jia, Q.Sun, B.Schiele, T.Tuytelaars, and L.Van Gool, Pose-guided person image generation, NIPS, 2017

[9] P. Esser, E. Sutter, and B. Ommer, A variational u-net for conditional appearance and shape generation, CVPR, 2018

[10] L.Zheng, Y.Yang, and A.G.Hauptmann, Person re-identification: Past, present and future, arXiv, 2016

[11] Z. Zheng, L. Zheng, and Y. Yang, A discriminatively learned CNN embedding for person reidentification, TOMCCAP, 2018

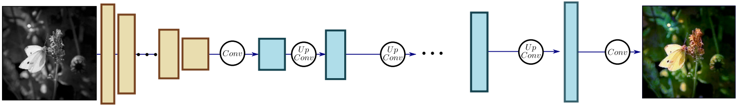
[12] A. Siarohin, **S. Lathuilière**, E. Sangineto, N. Sebe, Appearance and Pose-Conditioned Human Image Generation using Deformable GANs, TPAMI, 2019



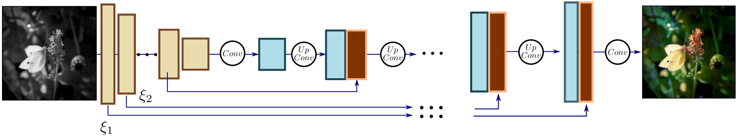
- Multiple input images
- How to select the relevant information in each image depending on:
    - pose difference
    - potential occlusions
    - image quality

[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020

U-Net [14]

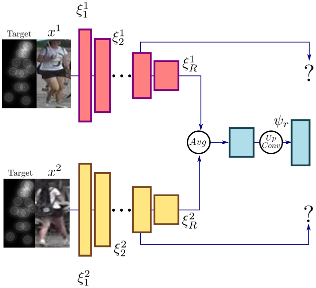


[14] O.Ronneberger, P.Fischer, and T.Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation

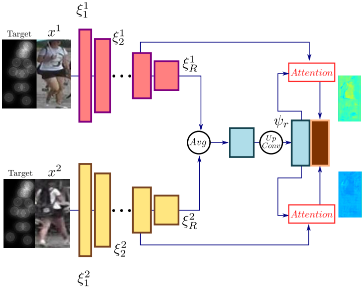


Skip connections

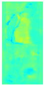
[14] O.Ronneberger, P.Fischer, and T.Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation



[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020

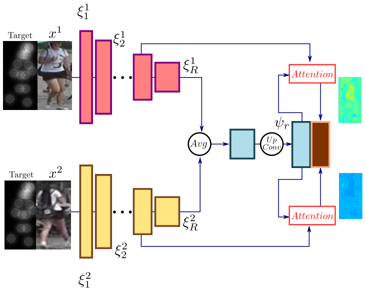


We propose [13]:



$$= \mathit{Att}(\underbrace{\psi_r}_{\text{features}}, \underbrace{\boldsymbol{\xi}_r^i}_{\text{skip connection}})$$

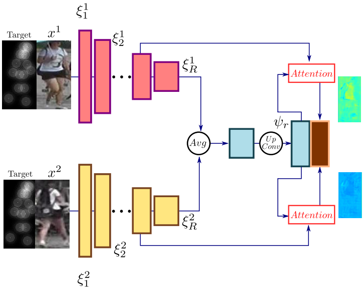
[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020



We propose [13]:

$$\color{red}{Att}\color{black}(\underbrace{\color{blue}{\psi_r}}_{\text{features}}, \underbrace{\boldsymbol{\xi}_r^i}_{\text{skip connection}}) \odot \boldsymbol{\xi}_r^i,$$

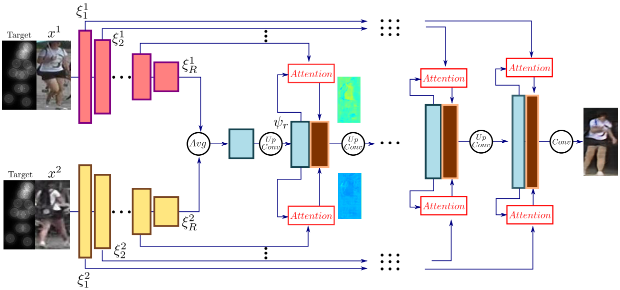
[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020



We propose [13]:

$$F_r = \sum_{i=1}^{M} Att(\underbrace{\psi_r}_{\text{features}}, \underbrace{\boldsymbol{\xi}_r^i}_{\text{skip connection}}) \odot \boldsymbol{\xi}_r^i, \tag{4}$$

[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020



We propose [13]:

$$F_r = \sum_{i=1}^{M} Att(\underbrace{\psi_r}_{\text{features}}, \underbrace{\boldsymbol{\xi}_r^i}_{\text{skip connection}}) \odot \boldsymbol{\xi}_r^i, \tag{4}$$

[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020

Results

Figure: A qualitative evaluation on the Market-1501 dataset.

[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020

Results

Figure: A qualitative evaluation on the Market-1501 dataset.

[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020

Figure: A qualitative evaluation on the Market-1501 dataset.

[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020

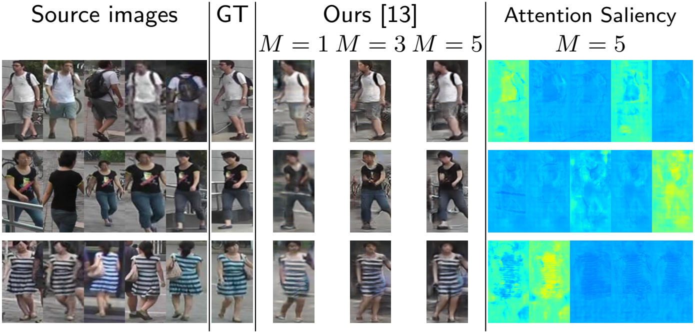
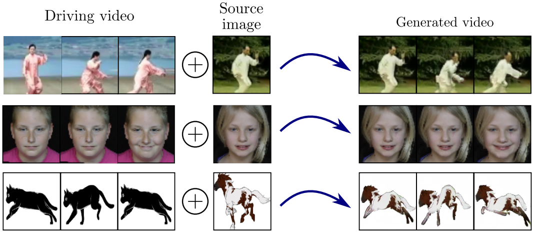
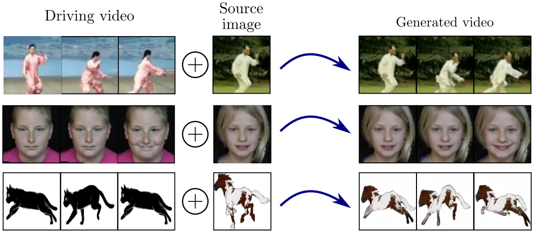


Figure: A qualitative evaluation on the Market-1501 dataset.

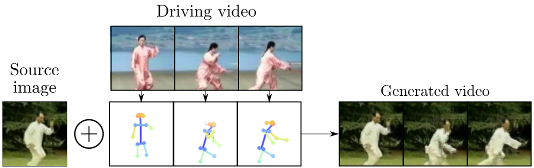
[13] **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020



21

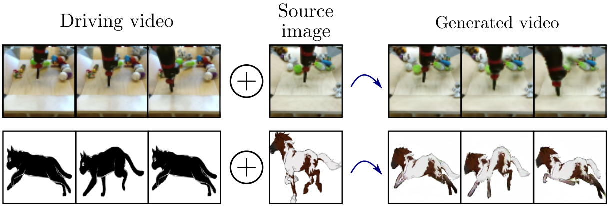


Naive solution: appearance transfer



Problems:

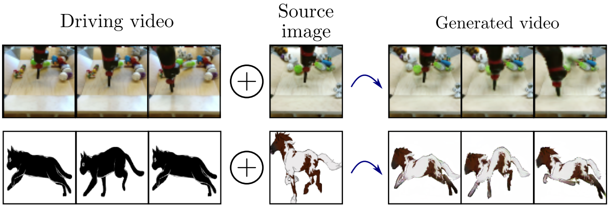
- It requires a detector



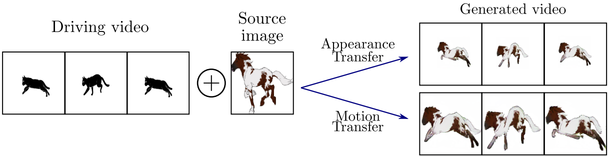
[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Problems:

- It requires a detector



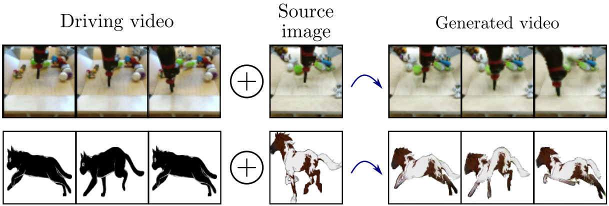
- Does not work when the shapes of the object are different



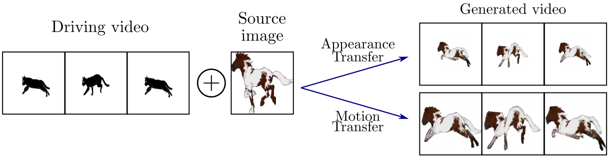
[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Problems:

- It requires a detector

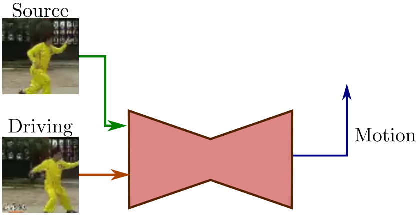


- Does not work when the shapes of the object are different



We propose: Self-supervised Motion Transfer [15].

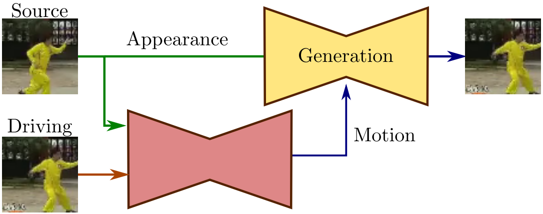
[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019



Self-supervised training.

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Self-supervised Motion Transfer [15]



Self-supervised training.

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

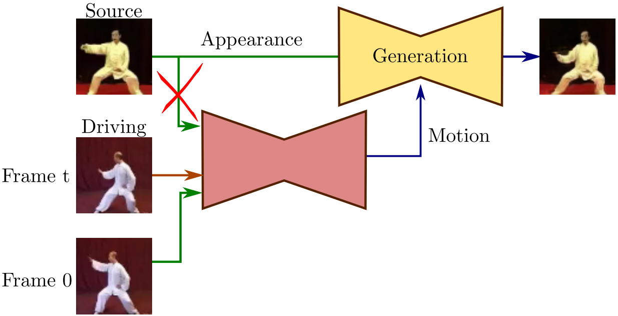
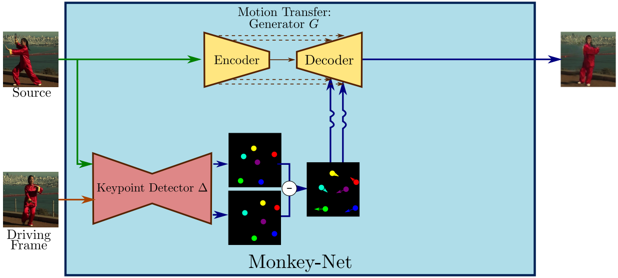
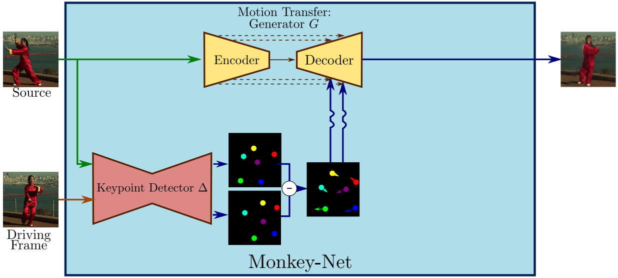


Image animation at test time.

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019



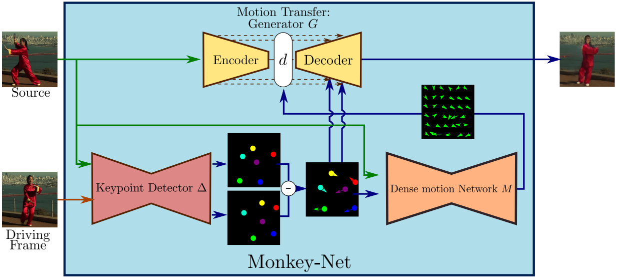
[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019



Again, we have an alignment problem.

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Monkey-Net: MOving KEYpoint Network [15]



Again, we have an alignment problem.

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

Image animation: Results

	Tai-Chi			Nemo (Face)			Bair
	$\mathcal{L}_1$	AKD	AED	$\mathcal{L}_1$	AKD	AED	$\mathcal{L}_1$
X2Face [16]	0.068	4.50	0.27	0.022	0.47	0.140	0.069
Ours [15]	**0.050**	**2.53**	**0.21**	**0.017**	**0.37**	**0.072**	**0.025**

Table: Video reconstruction comparisons. We employ *AKD: Average Keypoint Distance* and *AED: Average Euclidean Distance*

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

[16] X2Face: A network for controlling face generation by using images, audio, and pose codes. O.Wiles, A.S.Koepke, A. Zisserman, ECCV 2018

	Tai-Chi $\mathcal{L}_1$	*Tai-Chi* AKD	*Tai-Chi* AED	Nemo (Face) $\mathcal{L}_1$	Nemo (Face) AKD	Nemo (Face) AED	Bair $\mathcal{L}_1$
X2Face [16]	0.068	4.50	0.27	0.022	0.47	0.140	0.069
Ours [15]	**0.050**	**2.53**	**0.21**	**0.017**	**0.37**	**0.072**	**0.025**

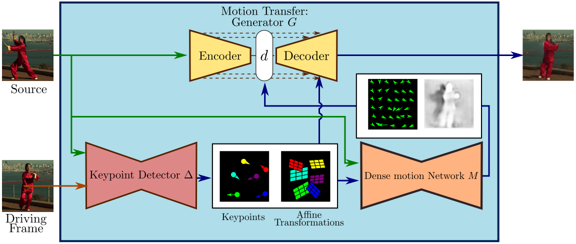
Table: Video reconstruction comparisons. We employ *AKD: Average Keypoint Distance* and *AED: Average Euclidean Distance*

Tai-Chi	*Nemo*	*Bair*
85.0%	79.2%	90.8%

Table: User study results on image animation. Proportion of times our approach is preferred over X2face [16].

[15] A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer , CVPR 2019

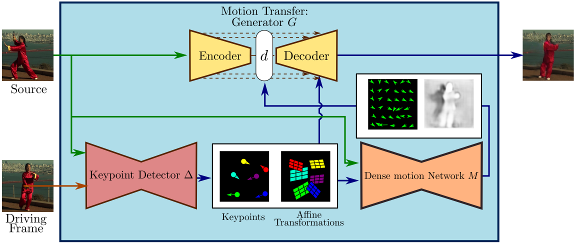
[16] X2Face: A network for controlling face generation by using images, audio, and pose codes. O.Wiles, A.S.Koepke, A. Zisserman, ECCV 2018



$$\mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p) = \mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p_k) + o(\|p - p_k\|), \tag{5}$$

[17] A. Siarohin, **S. Lathuilière**, S. Tulyakov, E. Ricci, N. Sebe, First Order Motion Model for Image Animation , NeurIPS 2019

Image animation: after Monkey-Net

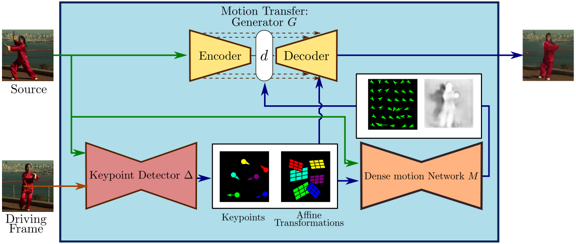


$$\mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p) = \mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p_k) + \left(\frac{d}{dp}\mathcal{T}_{\mathbf{X}\leftarrow\mathbf{R}}(p)\Big|_{p=p_k}\right)(p - p_k) + o(\|p - p_k\|), \quad (5)$$

[17] A. Siarohin, **S. Lathuilière**, S. Tulyakov, E. Ricci, N. Sebe, First Order Motion Model for Image Animation , NeurIPS 2019

Image animation: after Monket-Net

Image animation: after Monket-Net



Future works:

- Improve activity recognition methods
- Condition motion on other inputs
- Compression for video call (*e.g.* Skype)

Thank You!
Thanks to Aliaksandr, Sergey, Enver, Elisa and Nicu!

- A. Siarohin, E. Sangineto, **S. Lathuilière**, N. Sebe, Deformable GANs for Pose-based Human Image Generation, CVPR 2018
- A. Siarohin, **S. Lathuilière**, E. Sangineto, N. Sebe, Appearance and Pose-Conditioned Human Image Generation using Deformable GANs, T-PAMI, 2019.
- **S. Lathuilière**, A. Siarohin, E. Sangineto, and N. Sebe, Attention-based Fusion for Multi-source Human Image Generation, WACV 2020
- A. Siarohin, **S. Lathuilière**, S. Tulyakov , E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer, CVPR 2019
- A. Siarohin, **S. Lathuilière**, S. Tulyakov, E. Ricci, N. Sebe, First Order Motion Model for Image Animation , NeurIPS 2019