

## Today's outline

- ➊ Introduction [15 min] – Xavi
- ➋ Practical Deep Regression Guidelines [10 min] – Xavi
- ➌ Robust deep Regression with Probabilistic Models [20 min] – Xavi
- ➍ Introduction to Image and Video Generation [10 min] – Stéphane
- ➎ Pose-based Human Image Generation [15 min] – Stéphane
- ➏ Video generation: Image Animation [15 min] – Stéphane
- ➐ Conclusions [5 min] – Stéphane

Please do not hesitate to ask questions on the fly!!!

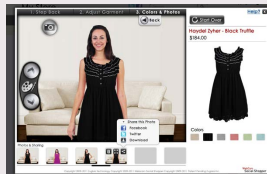
# Introduction to image and video generation

## Artistic / Editing / marketing purposes

- Photo editing



- Augmented reality <sup>a</sup>



- Many other applications: video games, animate face images from audio signal, increasing some intrinsic image properties...

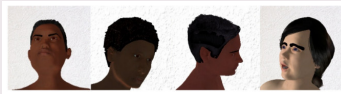
<sup>a</sup>image from *zugara* company

## Regression problems?

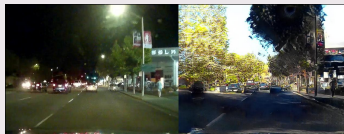
Regression with high dimensional target space.

## Machine Learning Tasks

- Generate annotated data: Head pose [1]



- Learning from few samples
- Domain adaptation

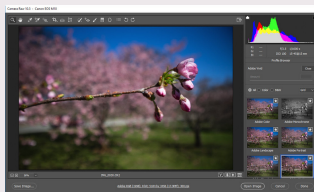


---

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

# How to generate images?

## Computer Graphics / Photo editing



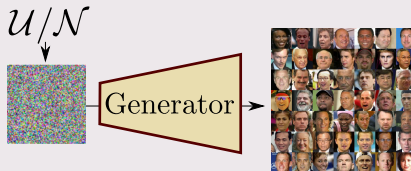
## Machine learning for image generation (since 2014)

- Deep generative methods
  - Generative Adversarial Network (GAN) [2]
  - Variational Autoencoder (VAE) [3]

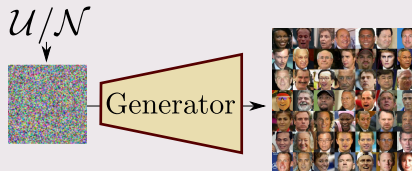
[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio., Generative adversarial nets, NIPS 2014

[3] D.P. Kingma, M. Welling, Auto-Encoding Variational Bayes, ICLR 2014

## Unconditional Image Generation



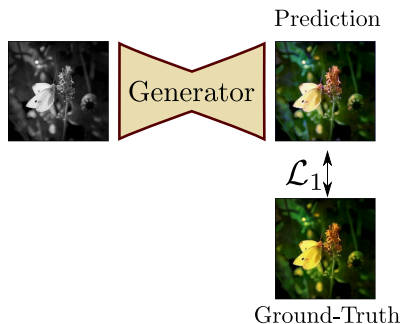
## Unconditional Image Generation



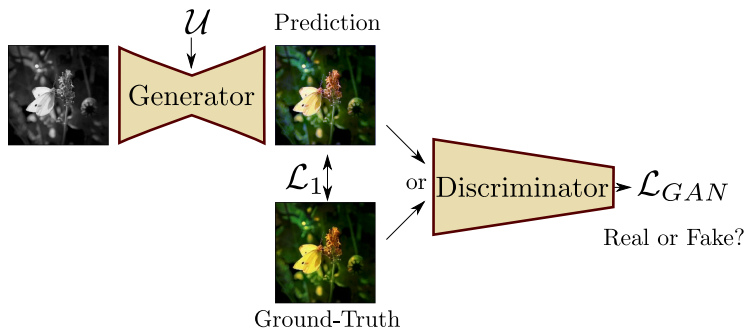
## Conditional: Image-to-Image translation [4]



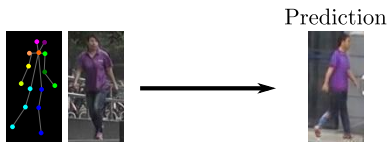
[4] P.Isola, J.-Y.Zhu, T.Zhou, A.A.Efros, Image-to-Image Translation with Conditional Adversarial Networks, CVPR, 2017

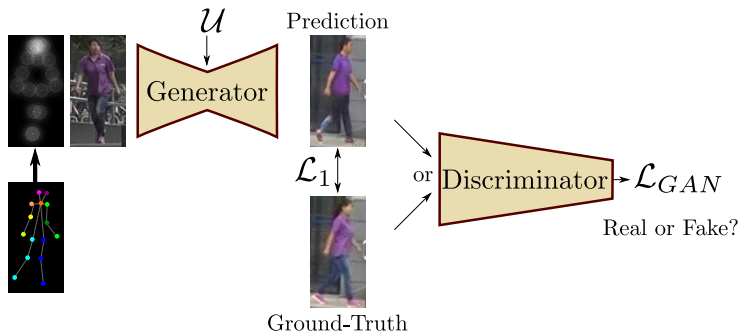




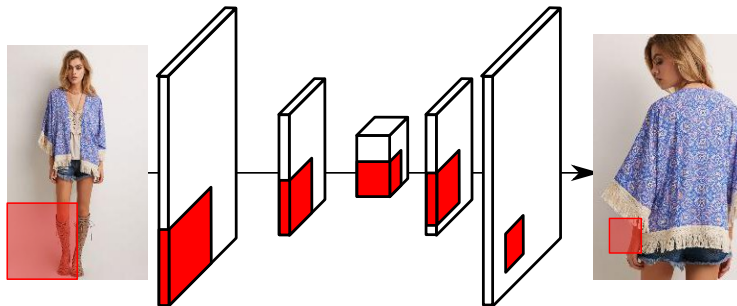


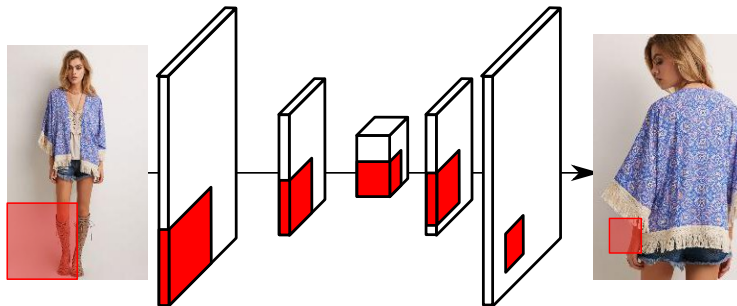
# Pose-based Human Image Generation





# Pose-based Human Image Generation



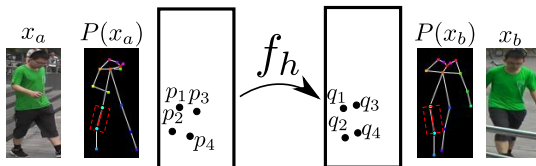


We need a deformation model!

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

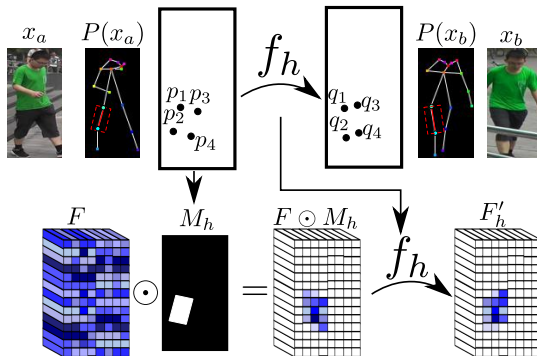


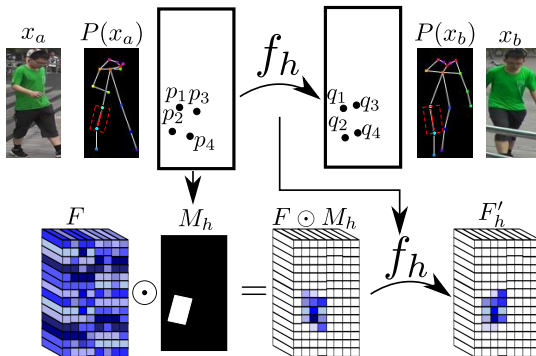
# Pose-based Human Image Generation: Our Proposal [6]





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





The body parts are combined:

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

- $\mathcal{L}_1$  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.

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

- $\mathcal{L}_1$  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.

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

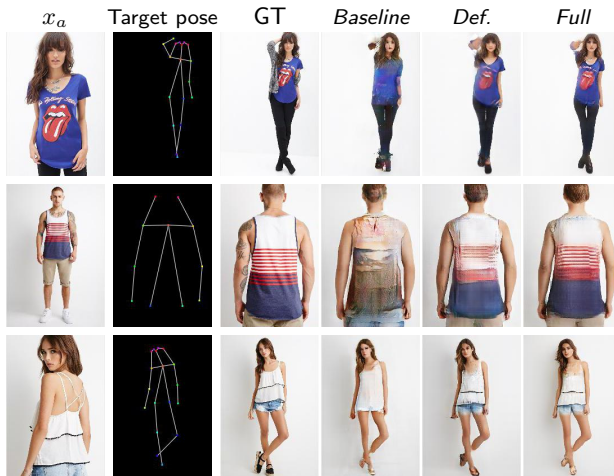
$$\mathcal{L}_{NN}(\hat{x}, x_b) = \sum_{\mathbf{p} \in g(\hat{x})} \min_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} \|g(\hat{x})(\mathbf{p}) - g(x_b)(\mathbf{q})\|_1, \quad (3)$$

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



**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	<i>SSIM</i>	<i>IS</i>
Ma et al. [7]	0.762	3.090
Ma et al. [8]	0.614	3.228
Esser et al. [9]	<b>0.786</b>	3.087
<i>Ours</i>	0.756	<b>3.439</b>

[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] <i>Rank 1</i>	Discr. Embedding [11] <i>Rank 1</i>
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, arXiv:1905.00007, under review for TPAMI, 2019





	IDE + Euclidean [10] <i>Rank 1</i>	Discr. Embedding [11] <i>Rank 1</i>
No augmentation	73.9	78.3
<i>Ours (Full) [12]</i>	<b>78.9</b>	<b>81.4</b>

**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, arXiv:1905.00007, under review for TPAMI, 2019



	IDE + Euclidean [10] <i>Rank 1</i>	Discr. Embedding [11] <i>Rank 1</i>
No augmentation	73.9	78.3
<i>Ours (Full)</i> [12]	<b>78.9</b>	<b>81.4</b>
<i>Ours (Baseline)</i>	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

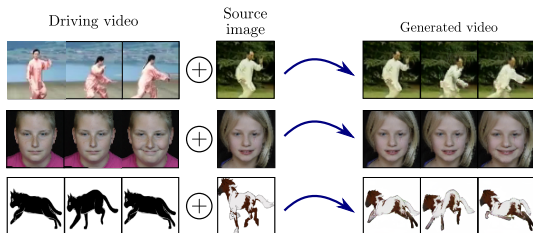
[10] L.Zheng, Y.Yang, and A.G.Hauptmann, Person re-identification: Past, present and future, arXiv, 2016

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

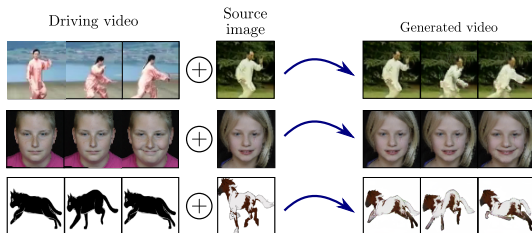
[12] A. Siarohin, **S. Lathuilière**, E. Sangineto, N. Sebe, Appearance and Pose-Conditioned Human Image Generation using Deformable GANs, arXiv:1905.00007, under review for TPAMI, 2019

## Video generation: Image Animation

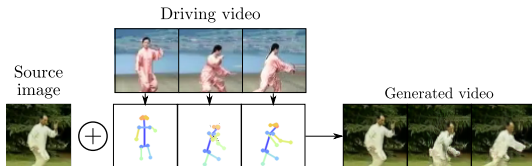
# Pose-guided generation for video generation?



# Pose-guided generation for video generation?

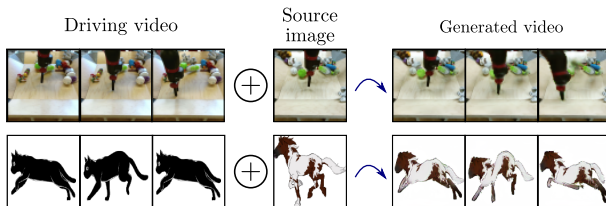


**Naive solution: appearance transfer**



## Problems:

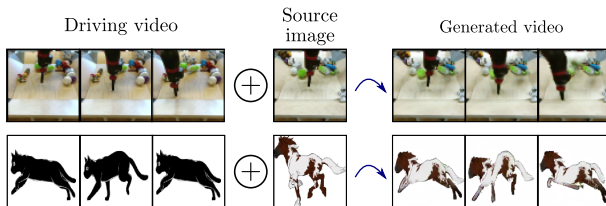
- It requires a detector



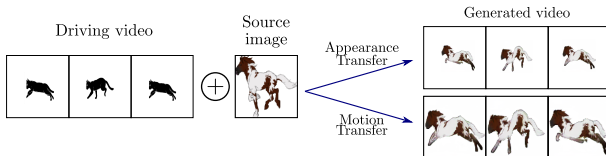
# Image animation: Appearance or Motion Transfer?

## Problems:

- It requires a detector



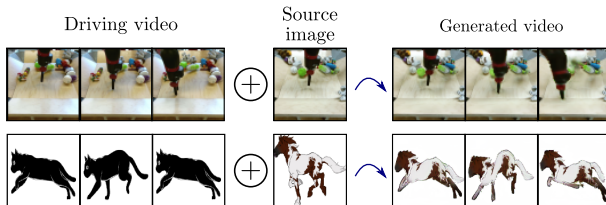
- Does not work when the shapes of the object are different



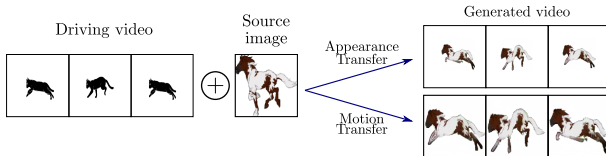
# Image animation: Appearance or Motion Transfer?

## Problems:

- It requires a detector



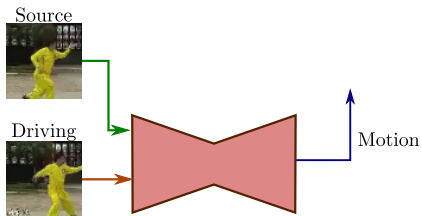
- Does not work when the shapes of the object are different



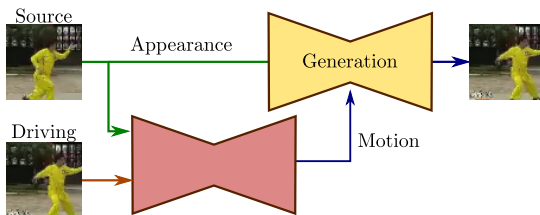
**We propose: Self-supervised Motion Transfer [15].**



# Self-supervised Motion Transfer [15]



Self-supervised training.



Self-supervised training.

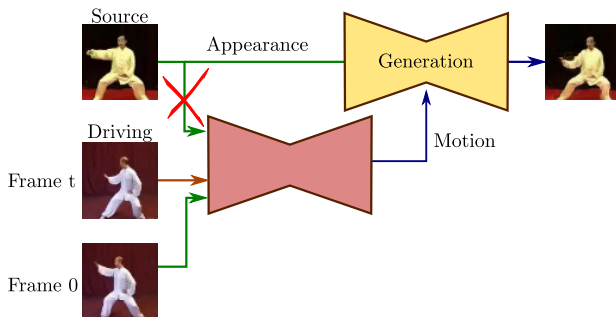
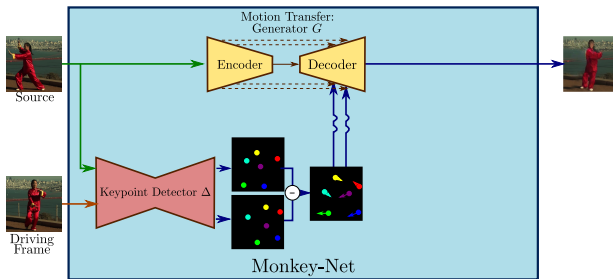
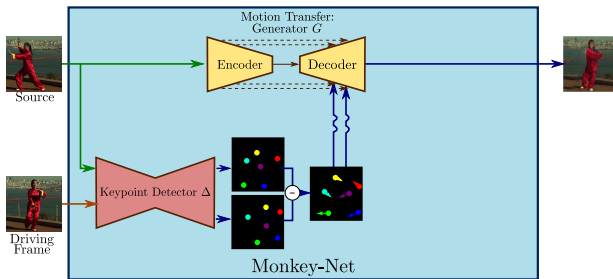


Image animation at test time.

# Monkey-Net: MOving KEYpoint Network [15]

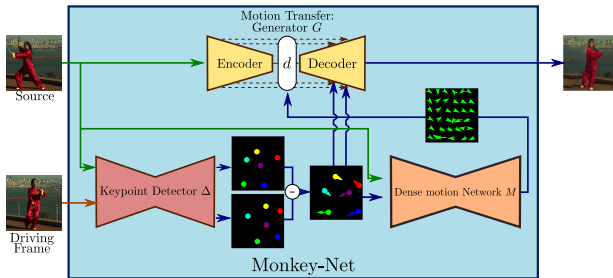


# Monkey-Net: MOving KEYpoint Network [15]



Again, we have an alignment problem.

# Monkey-Net: MOving KEYpoint Network [15]



Again, we have an alignment problem.









	<i>Tai-Chi</i>			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]	<b>0.050</b>	<b>2.53</b>	<b>0.21</b>	<b>0.017</b>	<b>0.37</b>	<b>0.072</b>	<b>0.025</b>

**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

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

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

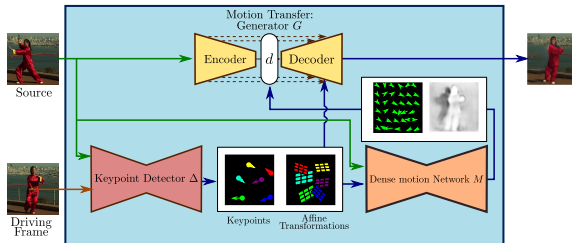
<i>Tai-Chi</i>	<i>Nemo</i>	<i>Bair</i>
85.0%	79.2%	90.8%

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

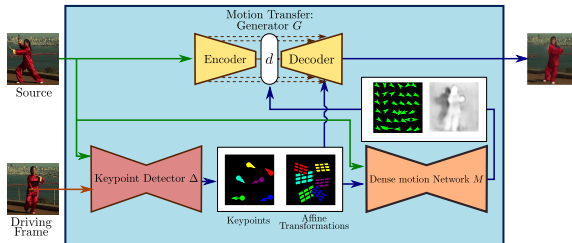
[15] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, Animating Arbitrary Objects via Deep Motion Transfer, CVPR 2019

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

# Image animation: after Monkey-Net



# Image animation: after Monkey-Net



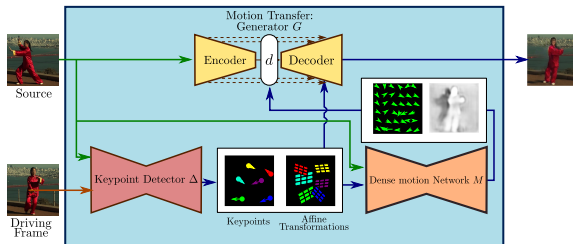
Video demo!











Video demo!

## Future works:

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

## Conclusion

We introduced deep regression as a general framework:

- Training deep architectures with a regression loss
- Use of pre-trained models
- Evaluation

We introduced deep regression as a general framework:

- Training deep architectures with a regression loss
- Use of pre-trained models
- Evaluation

## Question:

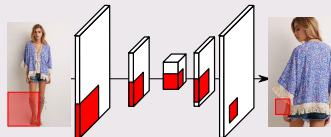
Is it really that general?

# Conclusion: 3 problems

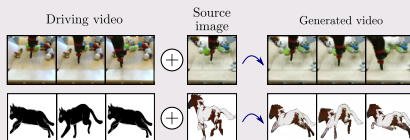
## Noisy training set



## Misalignment issue



## No supervision available



What about audio?



What about audio?

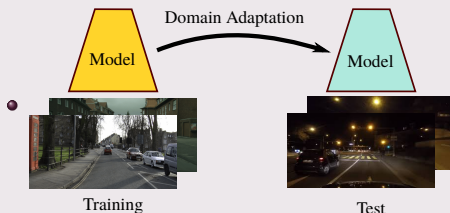


We should come back to our data and our problem!

## Many shallow regression models to *deepize*

- Regression with limited amount of data
- Robust regression

## Many unexplored directions for deep regression



- Lifelong learning
- Network architecture search



*Thank You!*

*Thanks to Aliaksandr, Yutong, Pablo,  
Sergey, Enver, Elisa, Radu and Nicu!*

- **S. Lathuilière**, P. Mesejo, **X. Alameda-Pineda**, and R. Horaud, DeepGUM : Learning Deep Robust Regression with a Gaussian-Uniform Mixture Model, ECCV 2018
- **S. Lathuilière**, **X. Alameda-Pineda**, P. Mesejo, and R. Horaud, A Comprehensive Analysis of Deep Regression, T-PAMI 2019
- 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, accepted with minor review at T-PAMI, 2019.
- 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