Image Analysis and Synthesis Deep Learning Use cases at technicolor

Louis Chevallier

Principal Scientist, Research and Innovation

technicolor.com

Teratec – Deep Learning workshop

20 Juin 2018

POWERING PREMIUM CONTENT ACROSS MARKETS

ADVERTISING

- Creative
- VFX
- Sound Finishing
- Color Finishing
- Immersive Experiences

FILM

- Dailies and Color pipeline management
- VFX
- Marketing Services
- Sound Finishing
- Color Finishing (including IMAX theatrical and HDR for home)
- DVD manufacturing and distribution

TELEVISION

- Dailies and Color pipeline management
- VFX
- Marketing Services
- Sound Finishing
- Color Finishing
- DVD manufacturing and Distribution

ANIMATION

- Original IP and production
- Asset creation
- Full servicing of film and television properties

GAMES

- Full servicing including asset and level building
- Sound Finishing
- Packaged media manufacturing and distribution

MUSIC

- Immersive Experiences
- Packaged media manufacturing and distribution

technicolor
FEEL THE WONDER

Impact of Deep Learning

Ackowledging outstanding performance of Deep Learning based solutions in computer vision.

- New functionalities emerge, requesting to revisit existing workflows.
- Higher performances calls for proper evaluation and metrics.
- Deep learning specific requirements raises integration and deployment challenges.

Uses cases

- Video Enhancement
  - *Upscaling*
  - denoising
- Video Editing, Augmentation
  - *Style Transfer*
  - *Mono to Stereo*
- Video encoding
  - *Compression*
- Asset Management
  - Indexing, Retrieval
  - *Classification*
- CGI, Animation
  - *Video 2 animation*

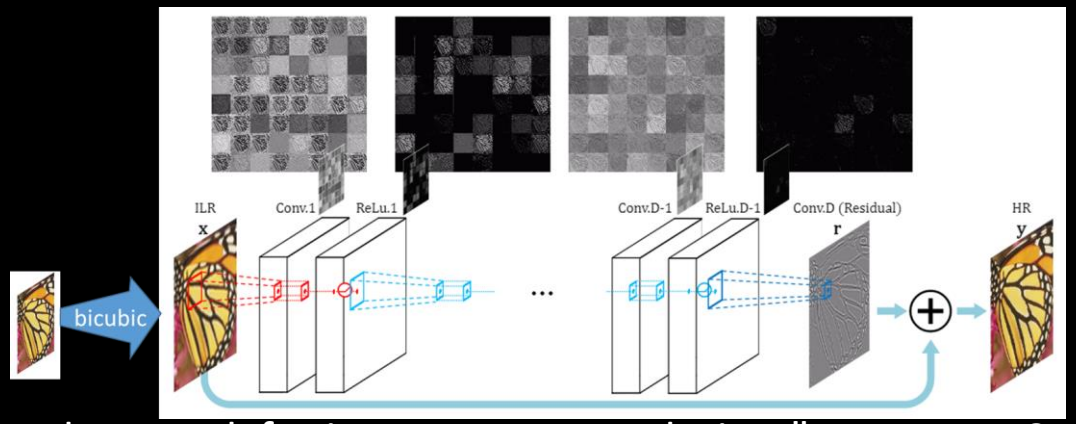
UC#1 Upscaling images

Converting images from 2K to 4K

X (HR) → Geometric Distortion → Blur, noise lens PSF → Subsampling Quantization → Y (LR)

Solution invert the distortions using a deep network : a stack of convolutionnal layers

Baseline



Dong, Chao, et al. "Learning a deep convolutional network for image super-resolution." *European Conference on Computer Vision*. Springer International Publishing, 2014.

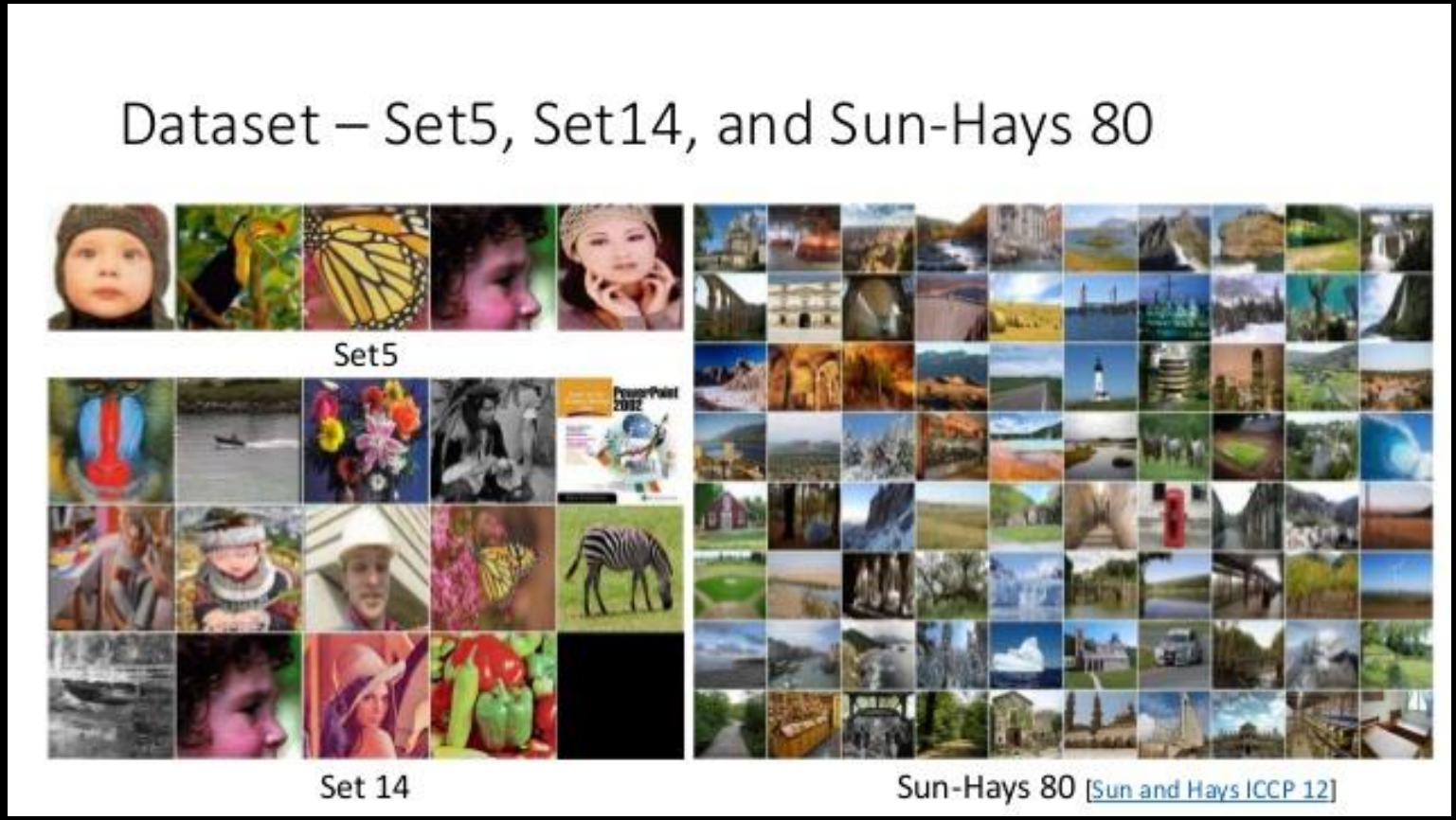
technicolor
FEEL THE WONDER

Evaluation

- Images captured at 2 different resolutions : LR, HR(x2)

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$



Accuracy – 2X scale factor:

	Bicubic	Deep
PSNR dB (Set5)	33.19	37.80

Speed :

HD (4K) image : about 1 sec with a GPU

technicolor
FEEL THE WONDER

Approach

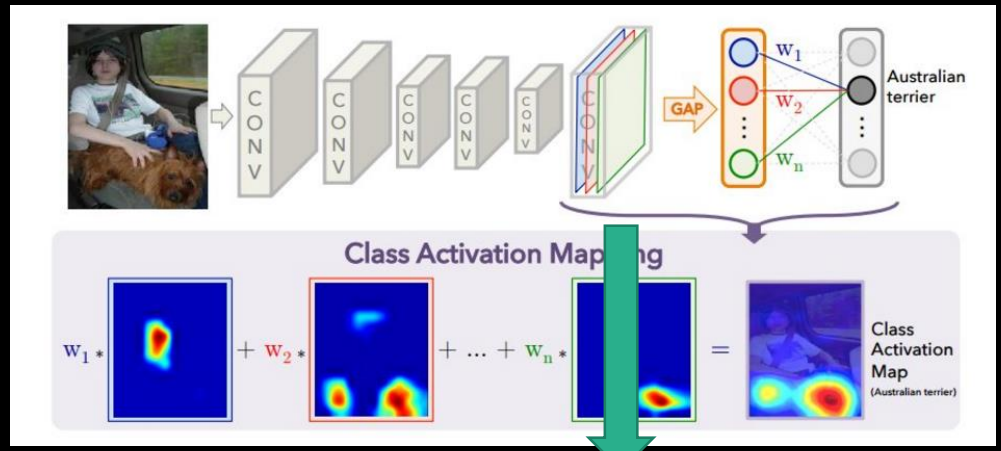
- Knowledge Transfer

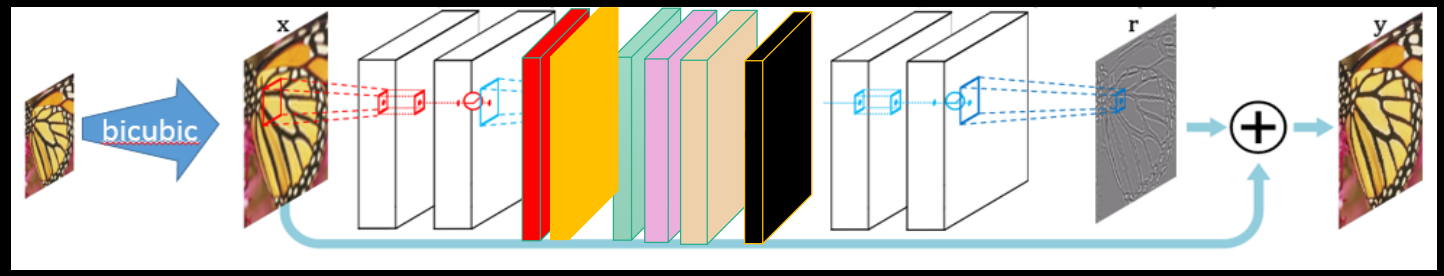
Filters learnt on ImageNet

Applying filters a, b conditionally

Y = a if K else b :

relu(a+K-1) + relu(b-K)

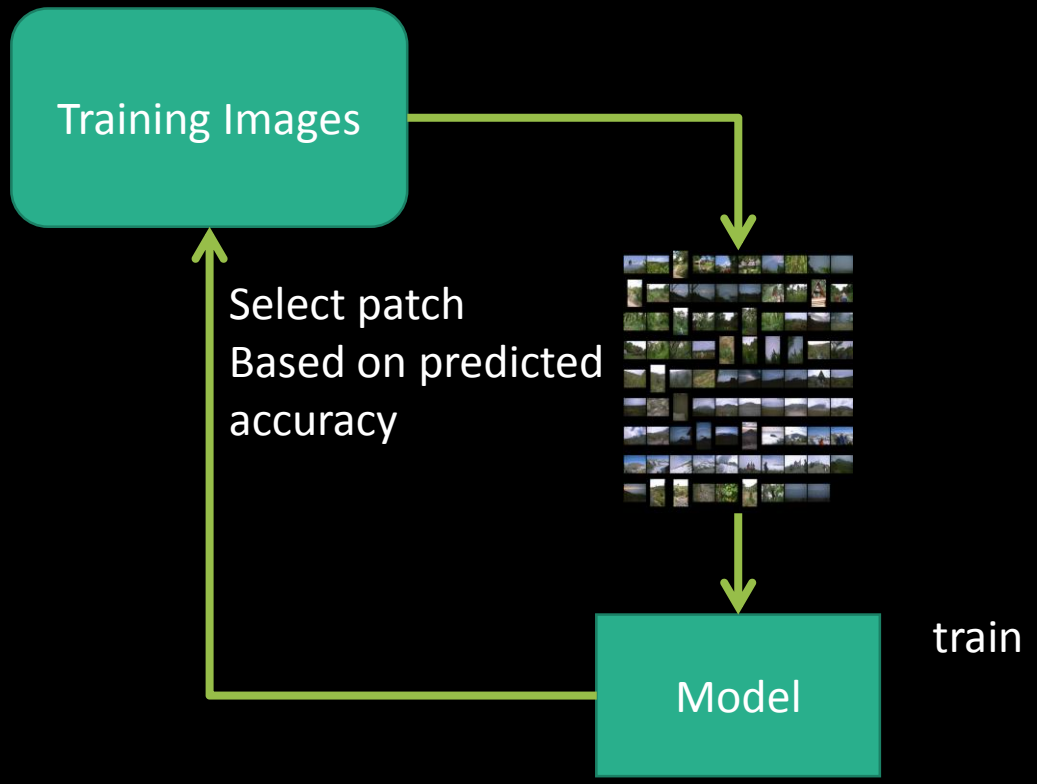




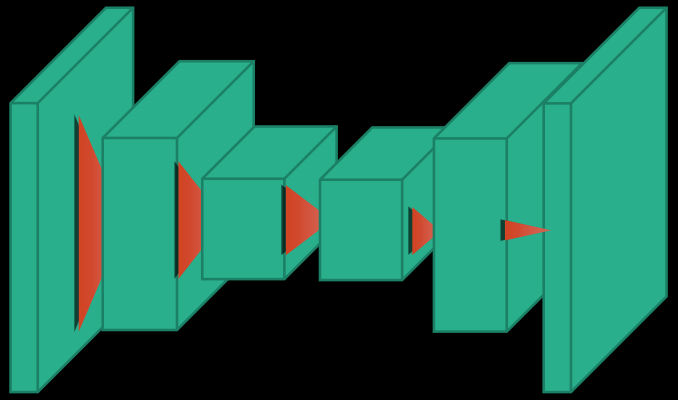
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921-2929).

Approach

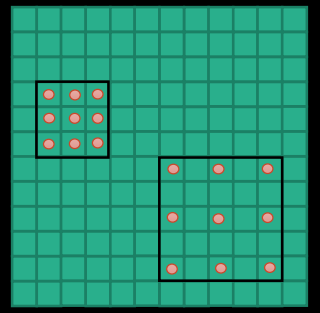
- Selective Sampling



- Wide receptive field capturing long range self similarities



Dilated convolutions



Ground truth




Deep approach




Standard approach

(Bicubic)

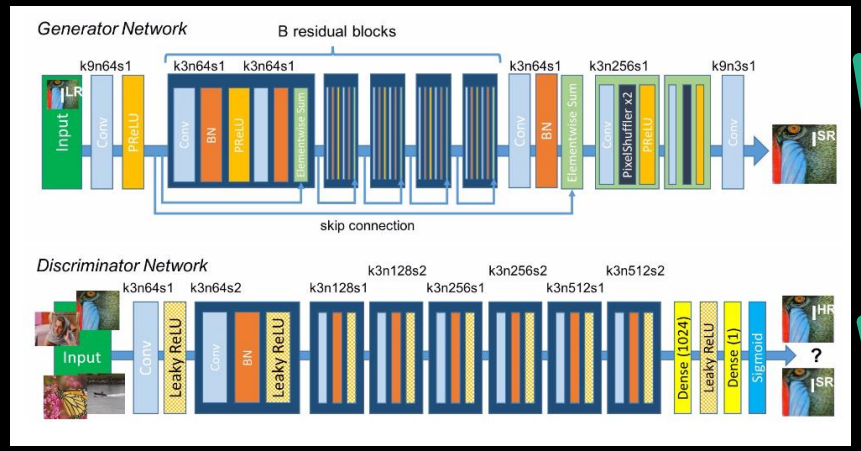



technicolor

FEEL THE WONDER

High order loss, GAN

« perceptual loss » != PSNR



Content Loss

Discriminator Loss

Bicubic DNN mse

GAN Original

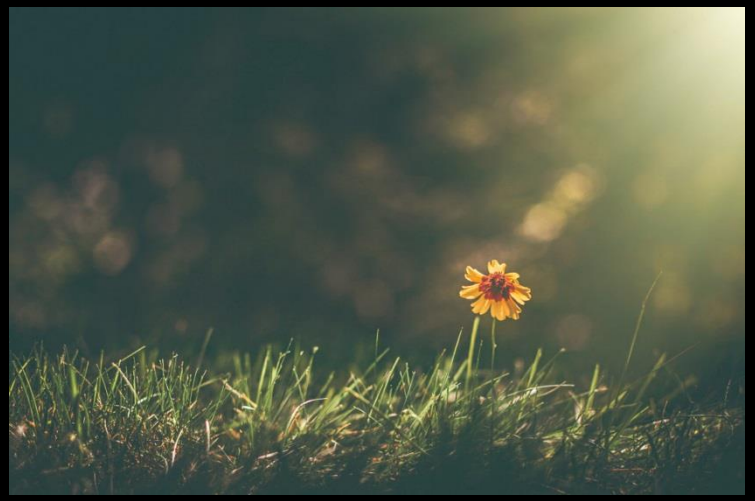
nice looking but PSNR no more applicable

Ledig, Christian, et al. "Photo-realistic single image super-resolution using a generative adversarial network." arXiv preprint arXiv:1609.04802 (2016).

UC#2 Predicting interestingness in Video and Images

- Which one is more interesting?

- Which one is more interesting?



technicolor

FEEL THE WONDER

Application

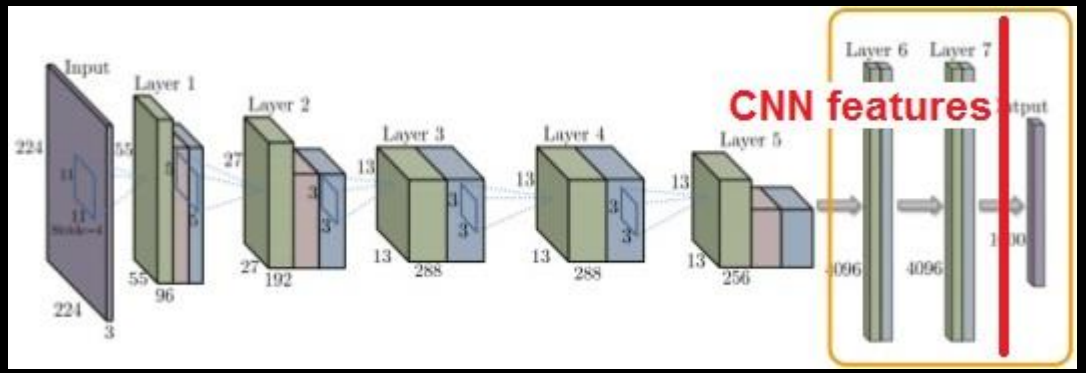
- Media file search & browse
- Advertisement

- Filtering and summarization
- E-learning

Approach

- From image / frame: CNN feature

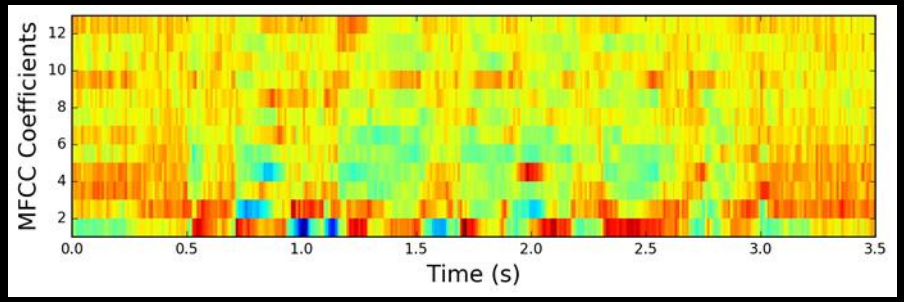
  output coefficients from a dense layer (fc7) of a CNN model (CaffeNet)
  Size = **4096**



- From audio: MFCC feature

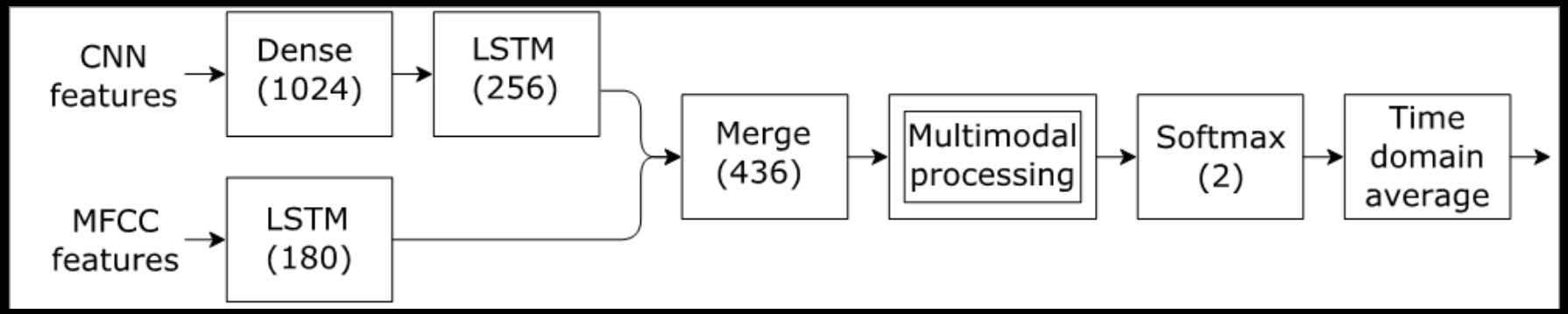
  Classic audio spectrum feature: Mel-frequency cepstral coefficients

  + Delta + Delta^2 , Size = 60 * 3 = **180**



technicolor
FEEL THE WONDER

Approach



Using LSTM layers + time domain average

Audio windows centered on frames for multimodal synchronization

Feature size handling and multimodal fusion

Results

Predicted interestingness: 0.00040983
Ground truth: not interesting

Predicted interestingness: 0.64466870
Ground truth: interesting

Evaluation

- Datasets
  - MediaEval (≈5000, unbalanced, human annotation)
  - Flickr (≈200000, balanced)

- Accuracy for Images, Mean Average Precision (MAP) metric for video
  A ranking based metric, used for MediaEval performance evaluation:

$$AveP = \sum_{k=1}^{n} p(k) \cdot \Delta r(k)$$

$$MAP = \frac{1}{m} \sum_{i=1}^{m} AveP_i$$

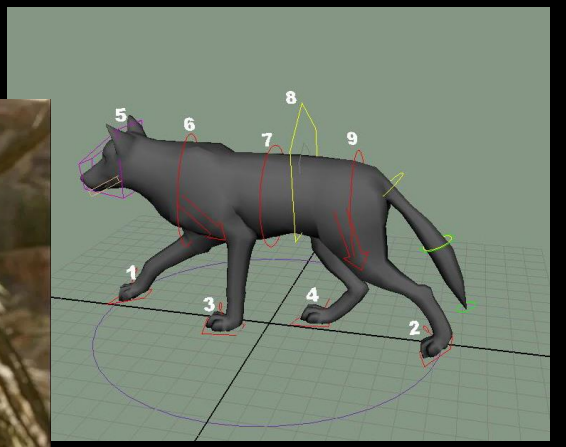
- Performance
  - Video : MAP = 0.37
  - Images + Textual metadata : accuracy = 97%

- Sufficient for targetted use cases : automatic video clip extraction

Adversarial content are possible

UC#3 Speeding up 3D Animation

Speeding up the production of animation movies

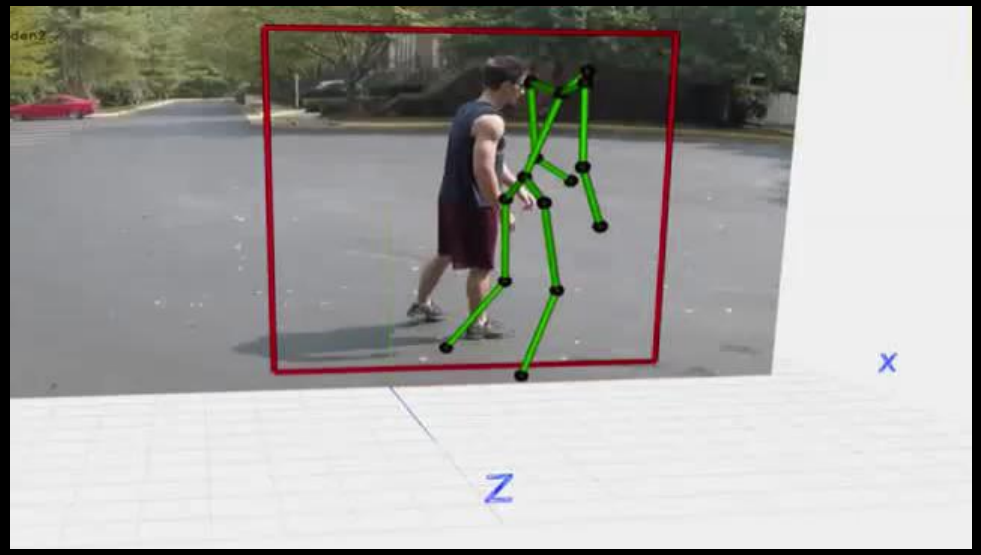


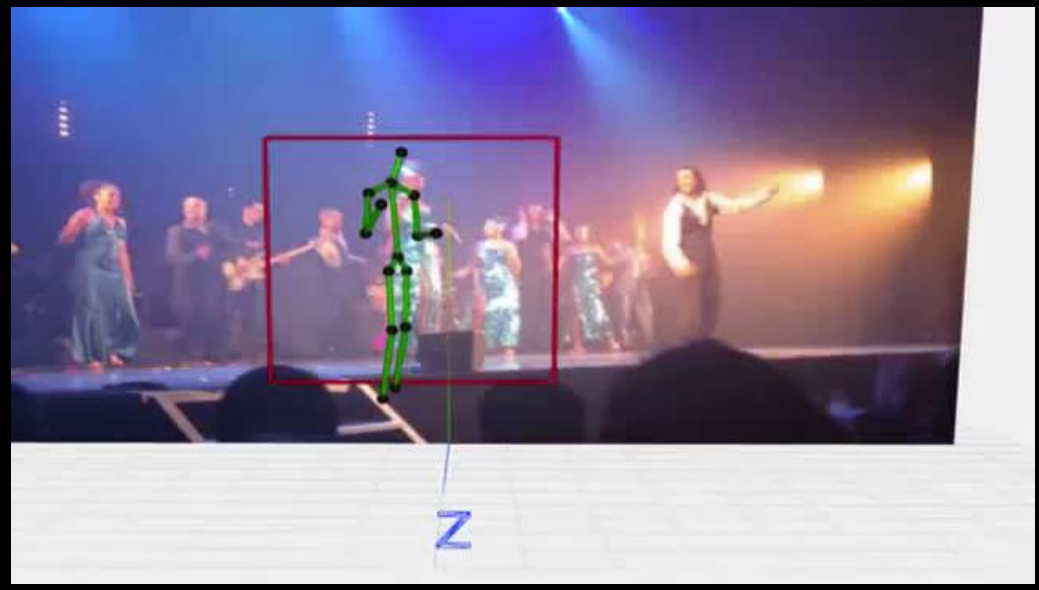
technicolor

FEEL THE WONDER

Video To Animation

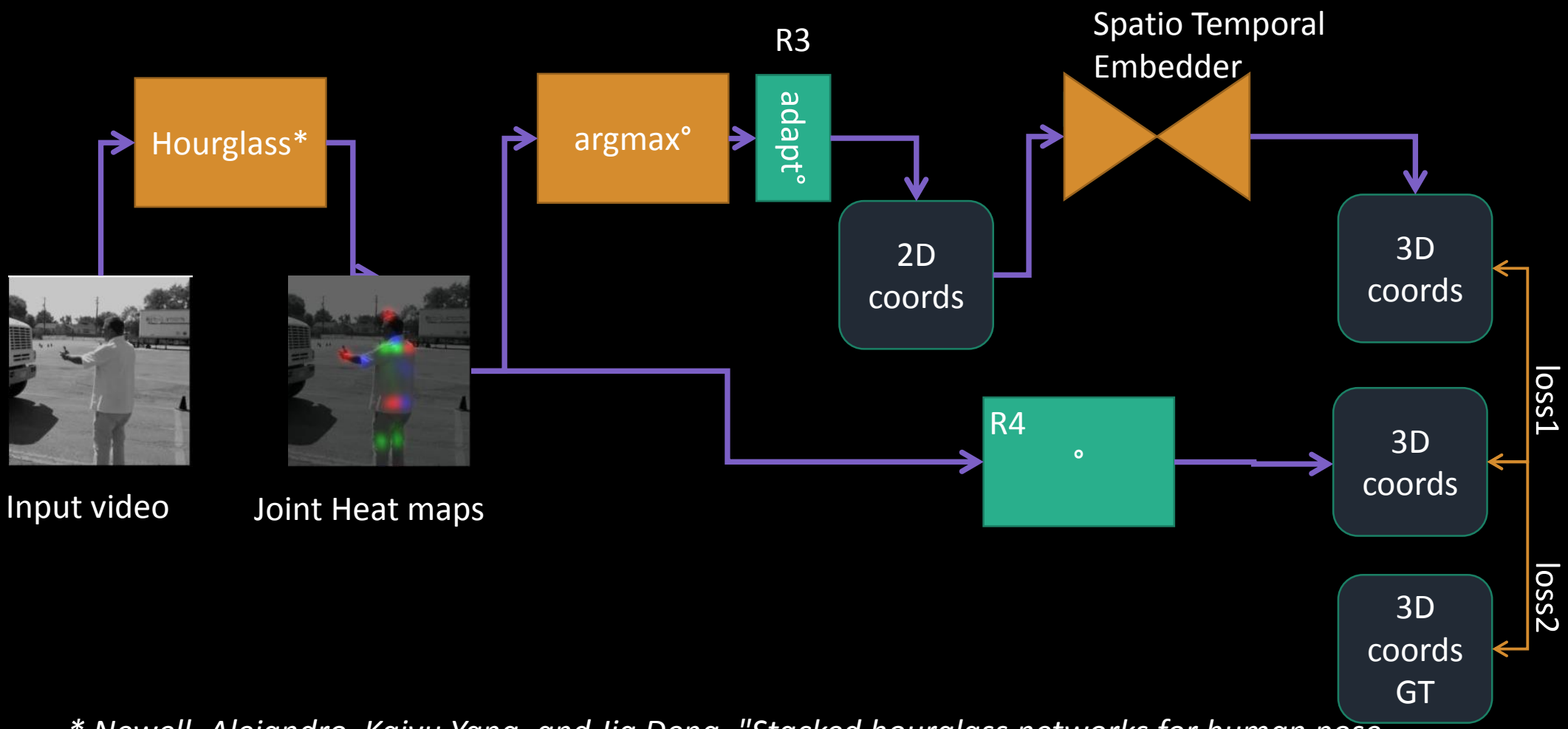
- Sketching animations starting from video





- Joints coordinates are extracted from images using plausible motions

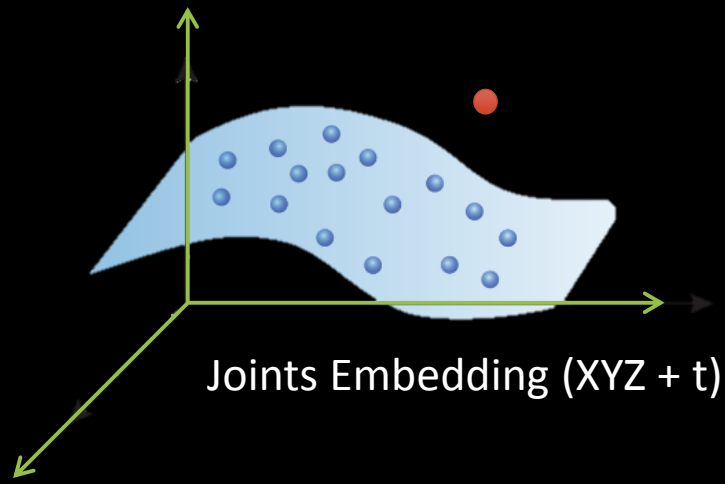
Approach

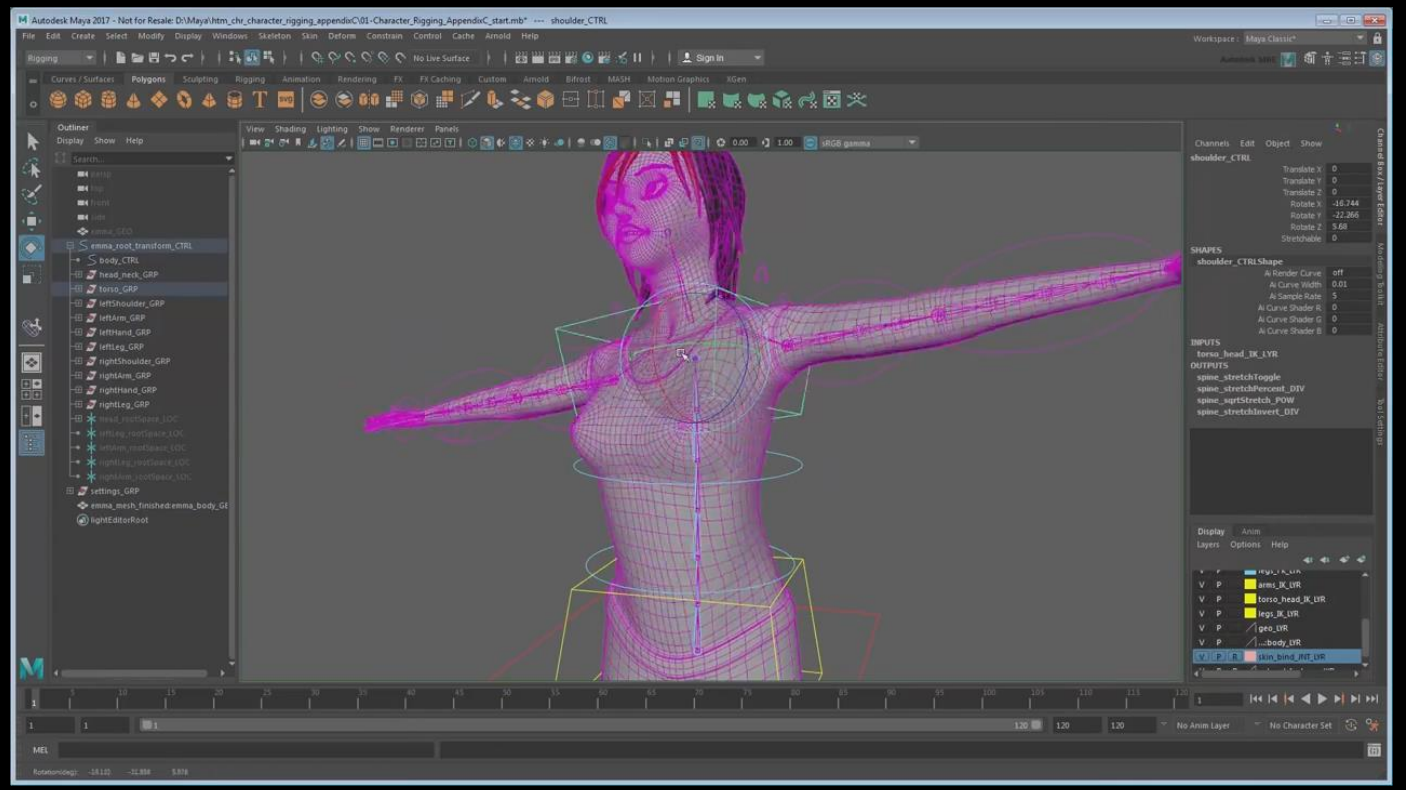


** Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European Conference on Computer Vision. Springer International Publishing, 2016.*

technicolor
FEEL THE WONDER

Integration





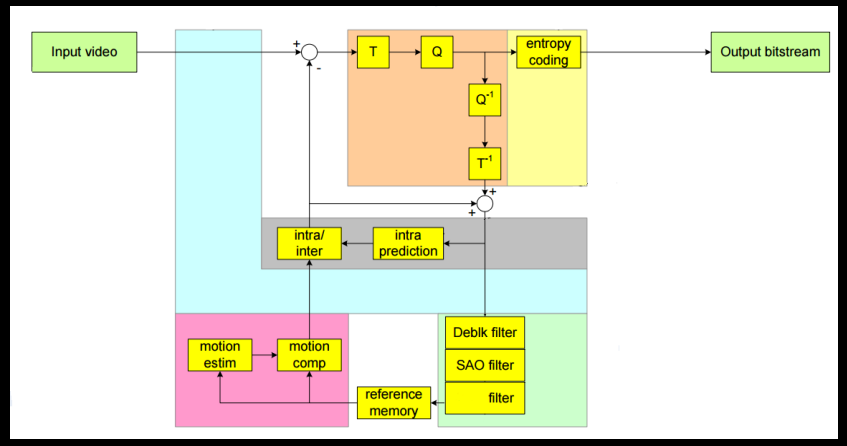
Animation are carried out by highly skilled artists with specialized GUI

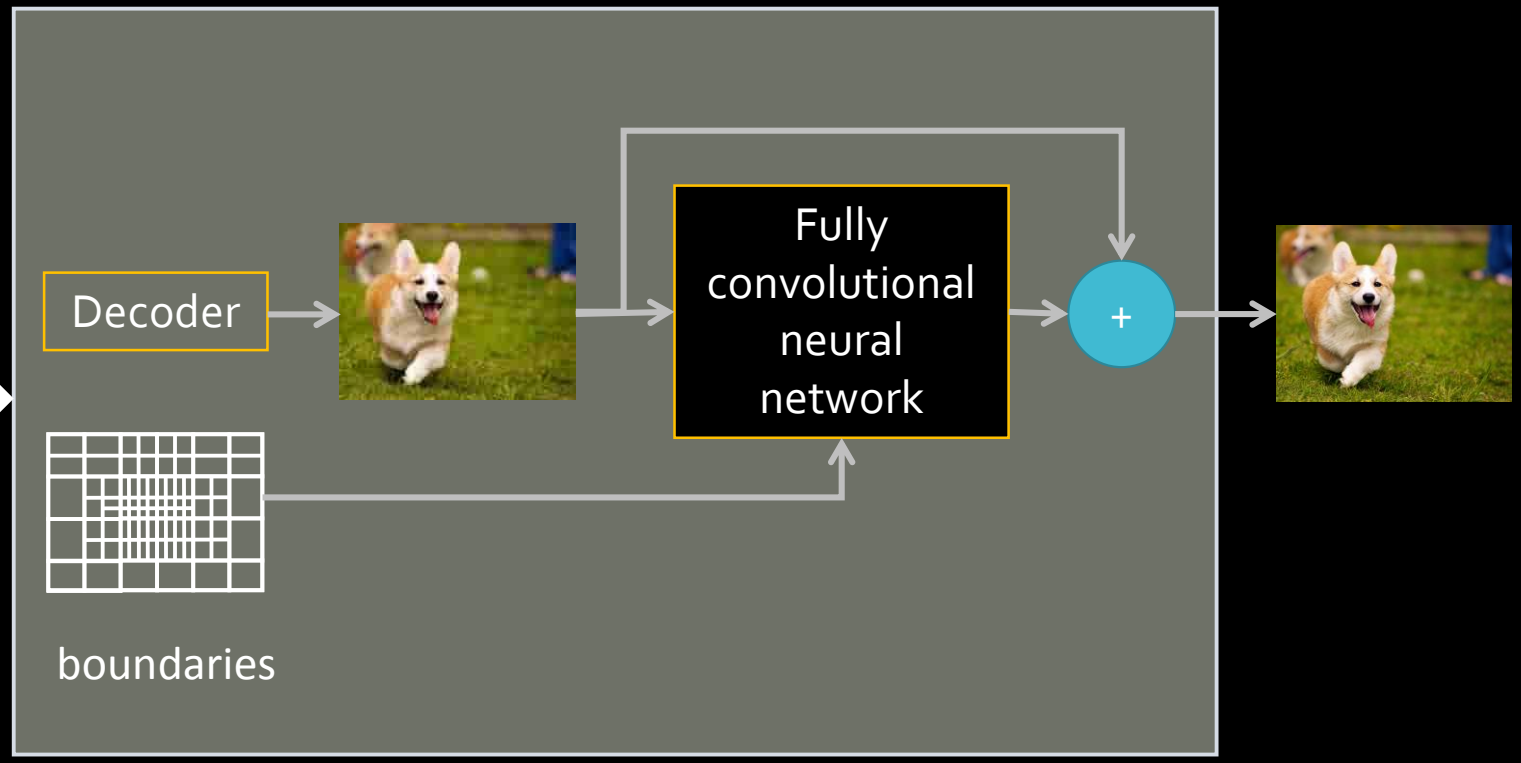
Mixing Rig controllers and learnt manifold compatibility

Need to devise new artist/machine interface

UC#4 Post-filters in future JEM video codec

Encoder



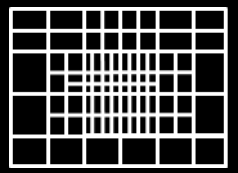


Dai, Yuanying, Dong Liu, and Feng Wu. "A convolutional neural network approach for post-processing in hevc intra coding." International Conference on Multimedia Modeling. Springer, Cham, 2017.

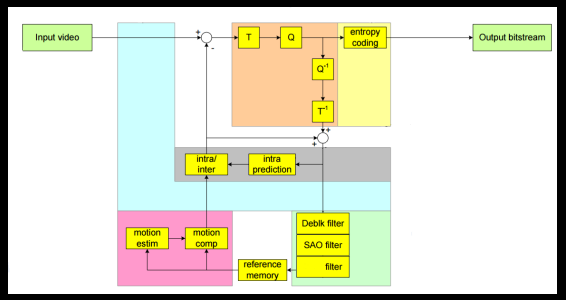
technicolor
FEEL THE WONDER

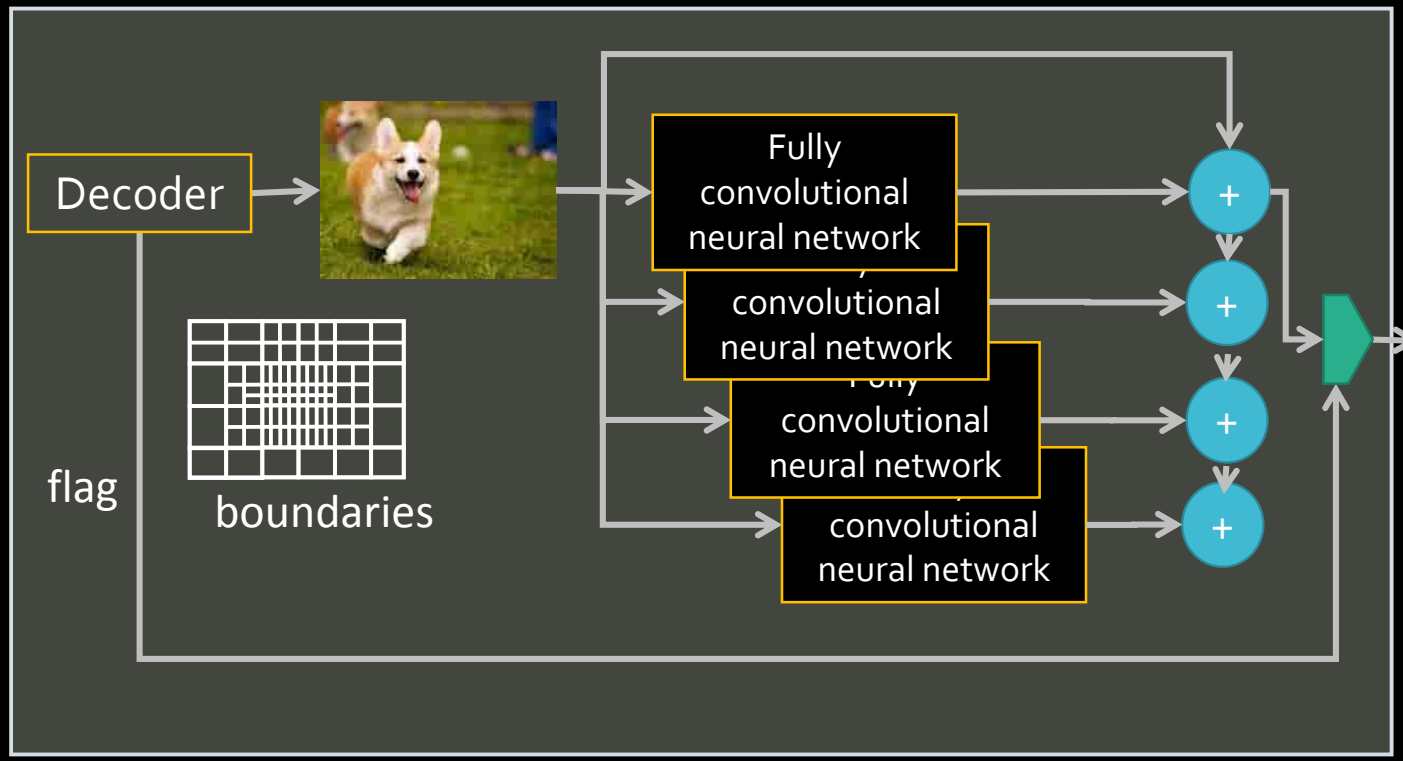
Approach

- Inform the network with boundaries



- Multibranch structure





technicolor
FEEL THE WONDER

Results

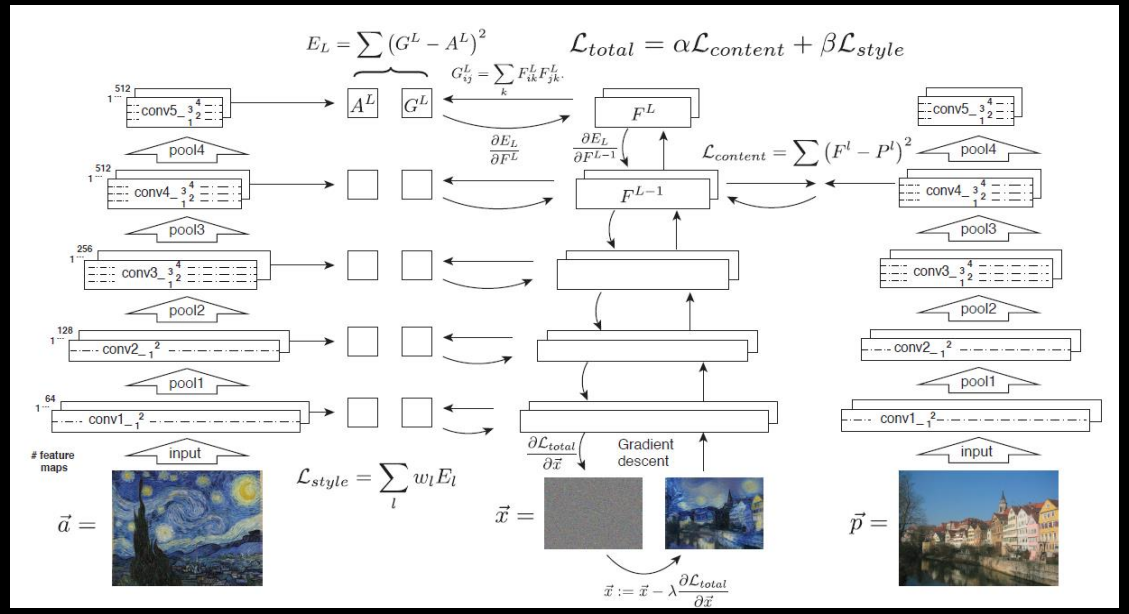
Post-filter	BD-rate
DBF + SAO + ALF	-3.2%
CNN	-4.91%

DNN has a high computation cost (4 layers, 30 000+ parameters)

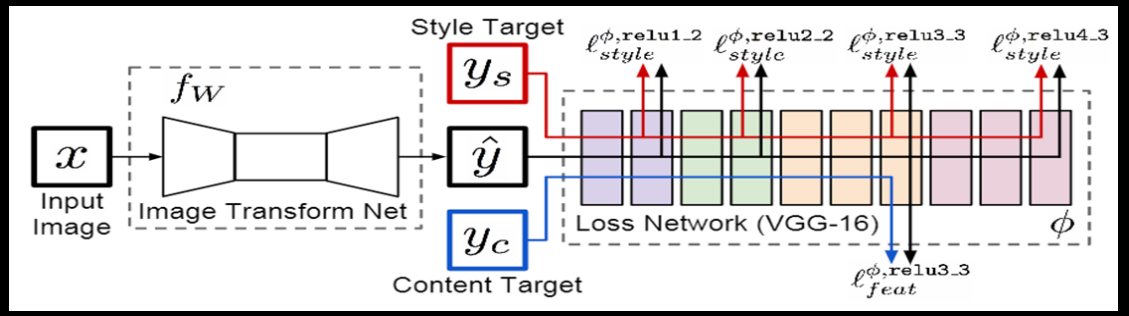
UC#5 Style Transfer

- A new editing tool

A neural Algorithm for Artistic Style



Fast feed forward network



Gatys et al.: "A Neural Algorithm of Artistic Style" 2015

Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." 2016.

technicolor
FEEL THE WONDER

Style transfer for AR

Style transfer for AR

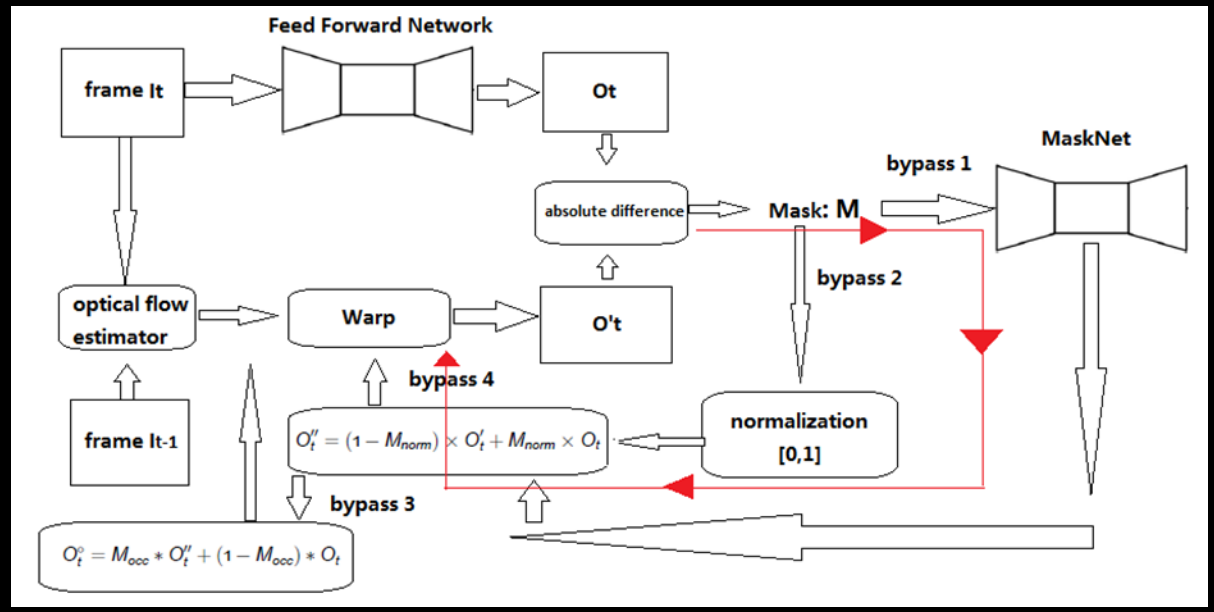
Fast Stable Style Transfer for Videos

- Real-time style transfer;

- Stability issue need to be adressed

Approach

- Use optical flow
  - Estimate t to t_{+1} discrepancy mask
- MaskNet to remove ghosting



Results

Speed up + 49% by reducing half of the convolution filters

+80% (temporal consistency loss) more stable

Original video

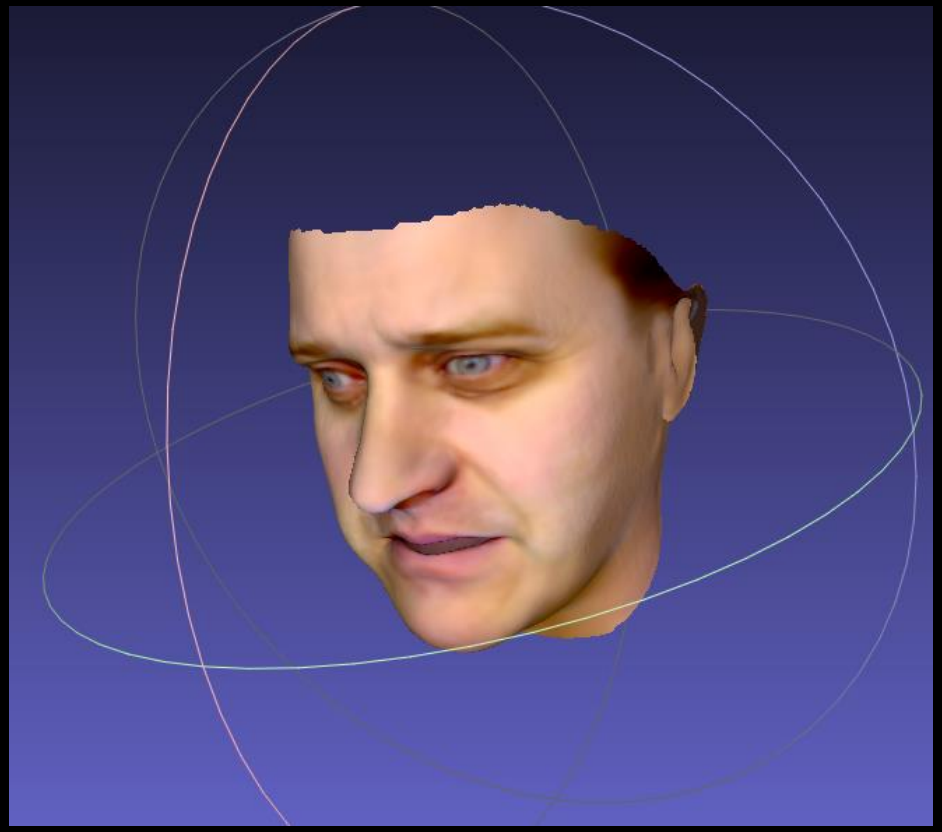
Johnson

simpleMask_fn2-ss

Now how to control the output, how to evaluate?

UC#6 3D modeling

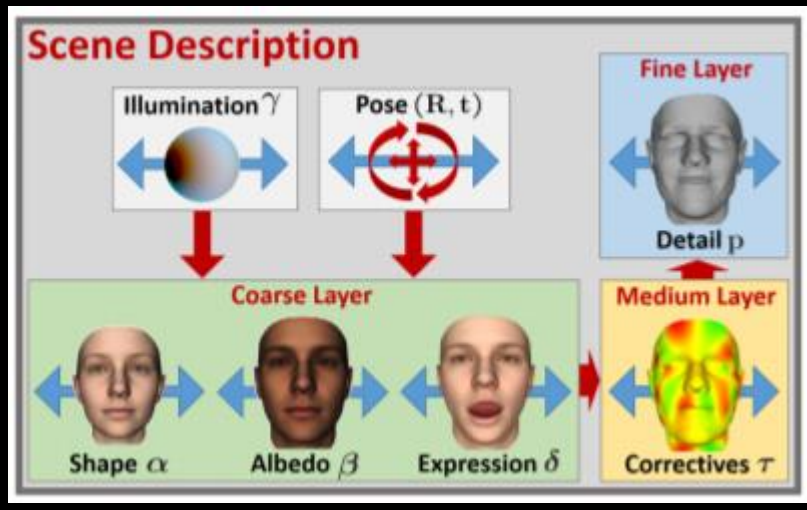
- Extracting facial rigs from monocular video
- High quality texture
- Dynamic wrinkle(exp)
- Face Edition

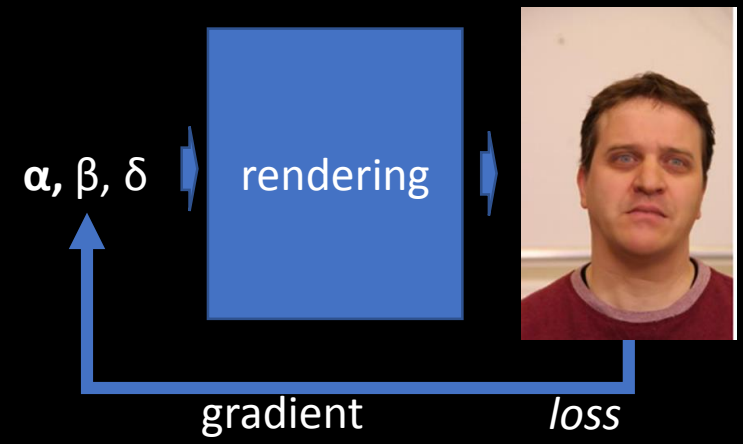


- Facial Reenactement, Complete Dubbing

Approach

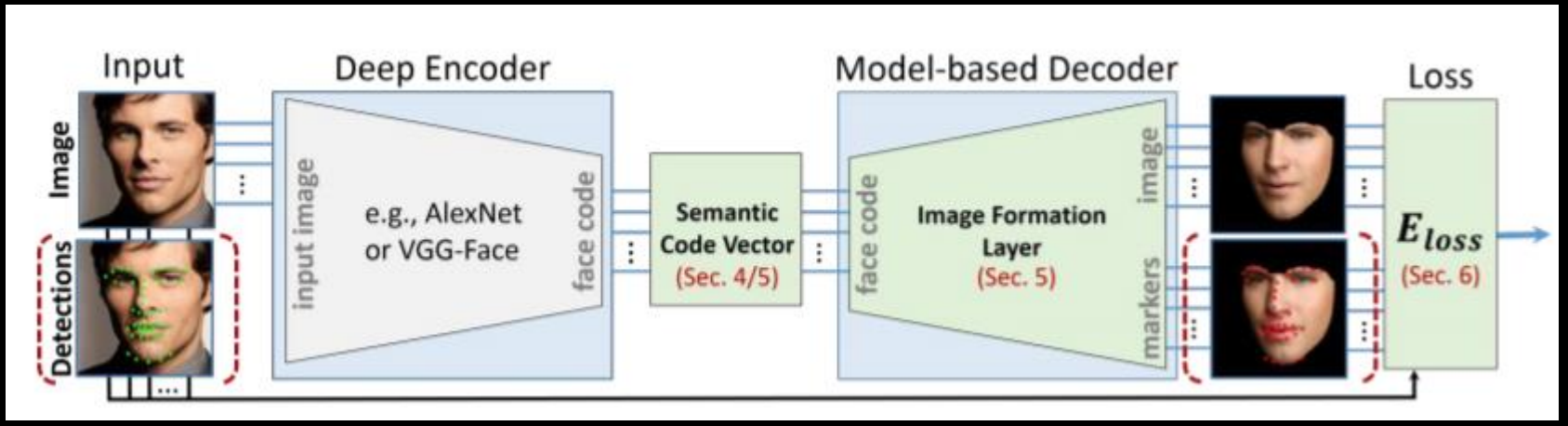
- From optimization of parameters (GD) to direct prediction
- Interactivity (Maya), speed vs quality





Optimization (slow, precise)

Direct prediction



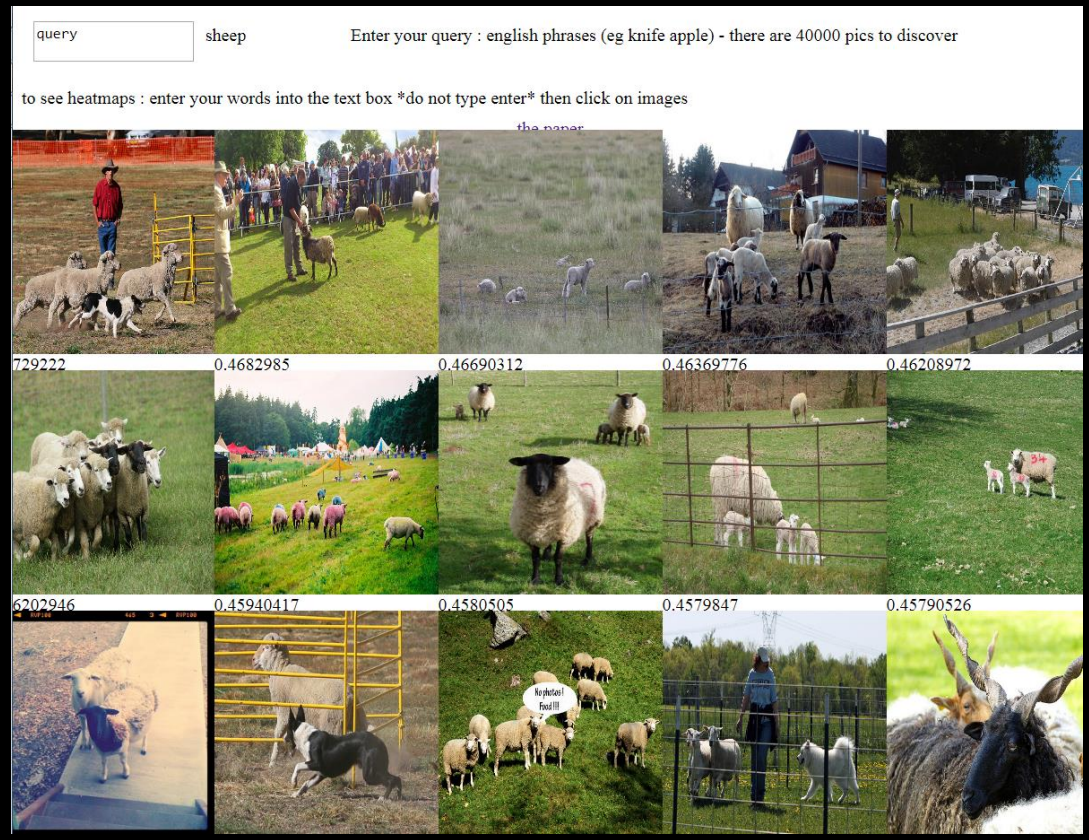
- Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., & Theobalt, C. (2017, October). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*

technicolor
FEEL THE WONDER

UC #7 Multimodal Text Video Embedding

- Retrieving images from natural text

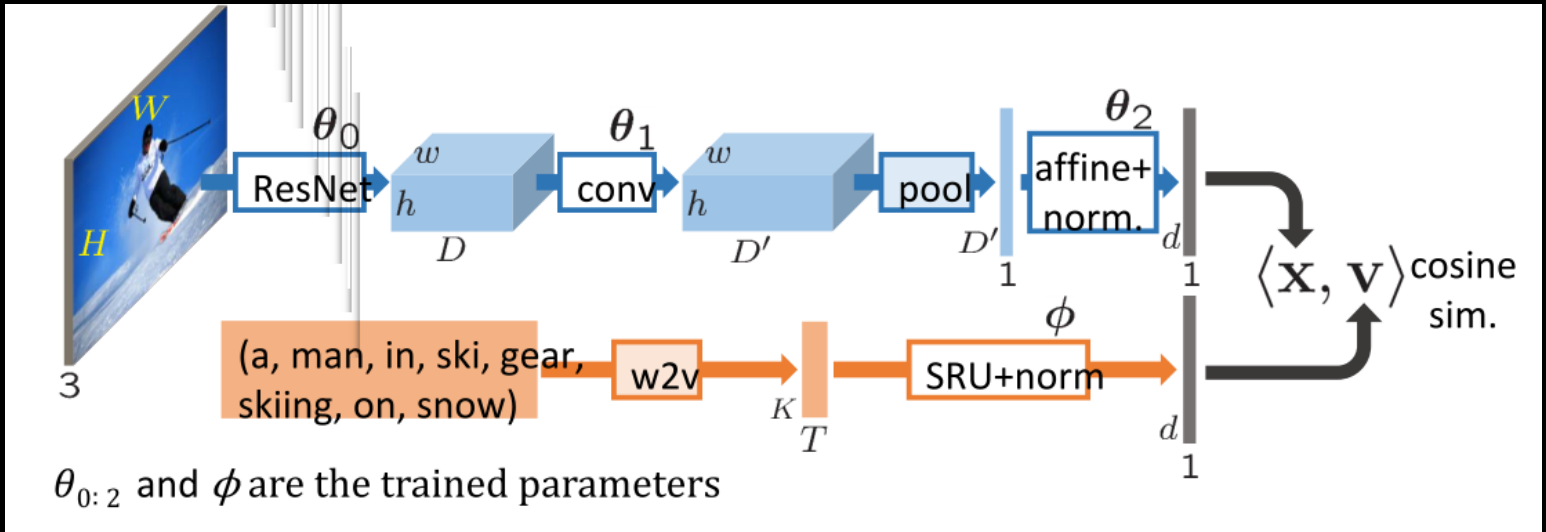
For an easy access to the huge data collected and generated in the course of movie production



technicolor
FEEL THE WONDER

Approach

- Visual pipeline
  - Resnet-152
  - Weldon spatial pooling
  - Affine Projection
- Textual Pipeline
  - Pre trained word embedding
  - SRU (rnn)



GT = textual annotation of images
Expansion/Specialization

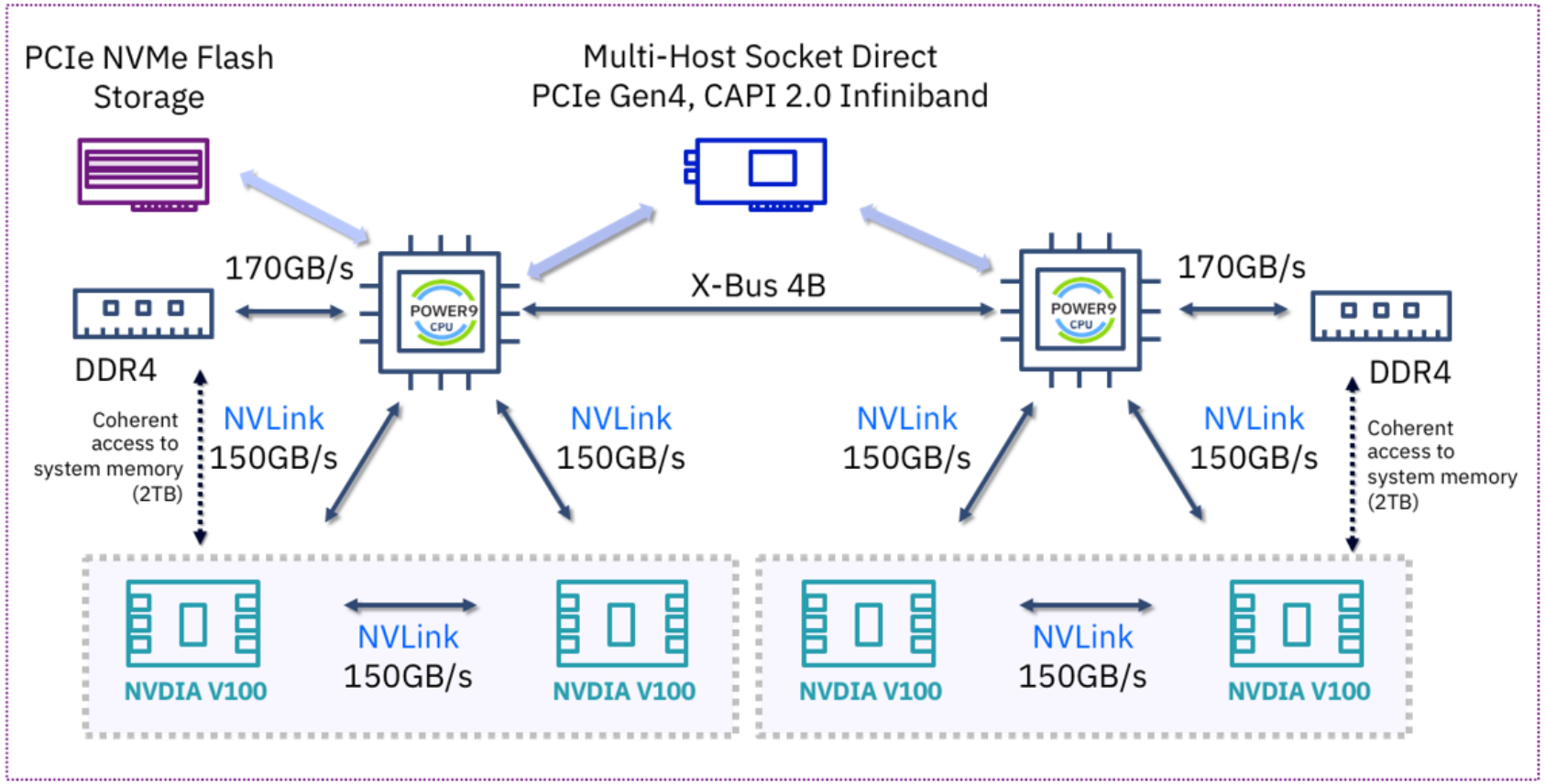
- ***Martin Engilberge, Louis Chevallier, Patrick Pérez, Matthieu Cord***; *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018*

technicolor
FEEL THE WONDER

- Positives and negatives pairs => Big batches on Multi-GPU with fast inter gpu transfer
- 3 days training on IBM Server using pytorch for joint use of several GPU

IBM AC922-GTG Server / Air-cooled

HW Architecture Overview



Conclusion

- DL allows substantial improvements in several applications
- Integration challenges need be solved
- Existing workflows have to be adapted
- Metrics are needed for new tools
- Stability, Interactivity needs further research
- Many new deep based tools are emerging and more are still to be discovered
- Many training experiments required : GPU indispensable.

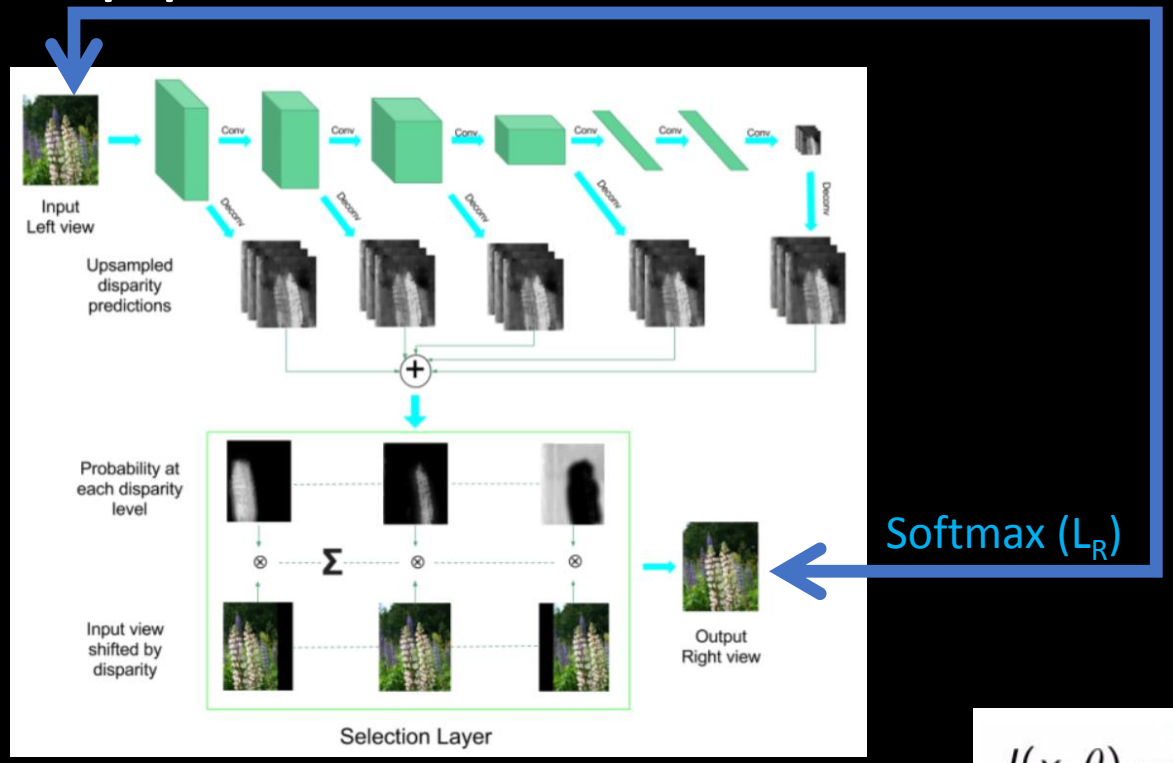
Thank You

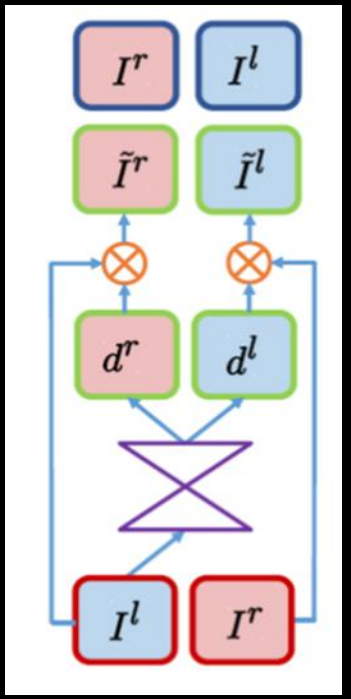
technicolor

FEEL THE WONDER

UC#8 Stereo Panoram for VR
Stereo from Mono

Approach





$$J(x,\theta) = \alpha_R L_R + \alpha_L L_L + \alpha_{TV} L_{TVL} + \alpha_{TV} L_{TVR} + \alpha_{LRC} L_{LRC} + \lambda L_{reg}$$

- *Xie, Girshick, Farhadi: Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural network (2016).*

Results