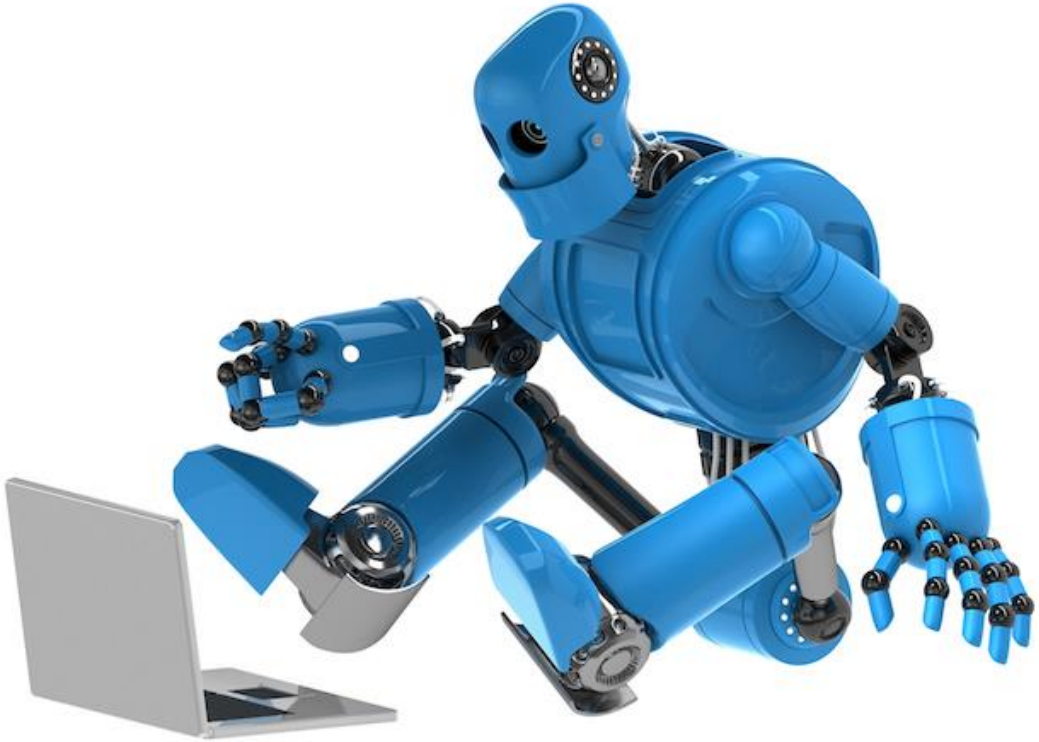
Deep Augmentation for image and activity recognition

By Abdullah Hamdi



outline

- Brief history (CNN , ImageNet, activity recognition)
- Brief history (DeconvNets , GAN, applications)
- My previous work on activity generation
- Recent GAN papers:
  - Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)
  - Adversarially Tuned Scene Generation (CVPR'17)
  - Generating Videos with Scene Dynamics (NIPS'16)
- What's next ?

Image classification (old days)

- Hand crafted features
- Bag of words

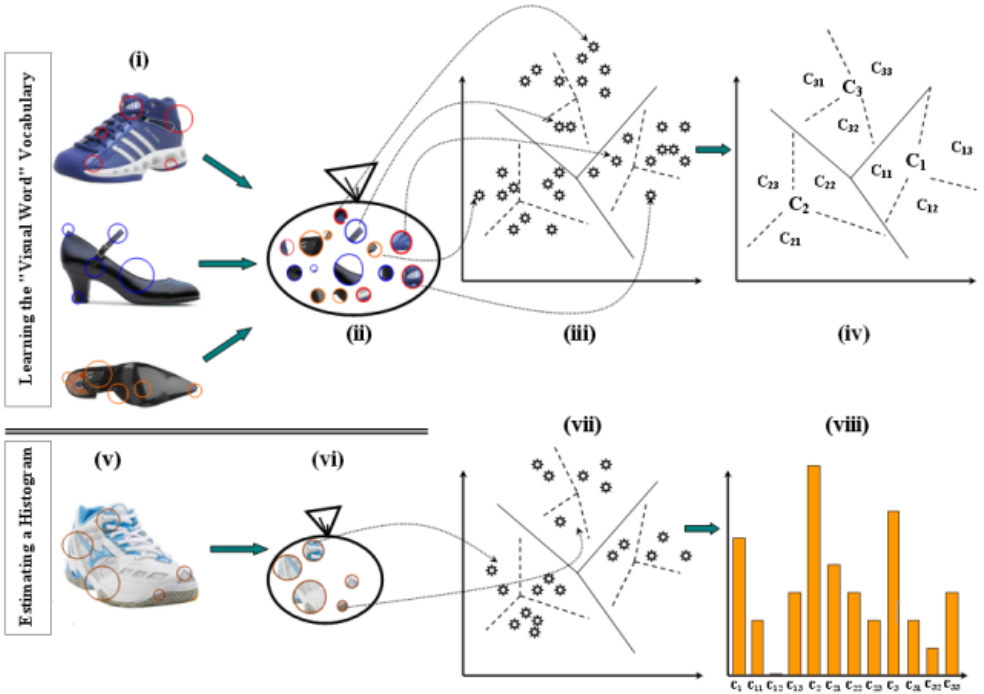
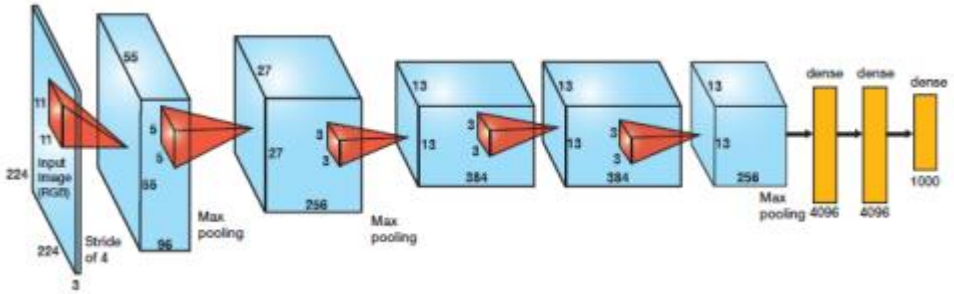


Image classification (new era)

- Deep CNNs
  - AlexNet
  - VGG



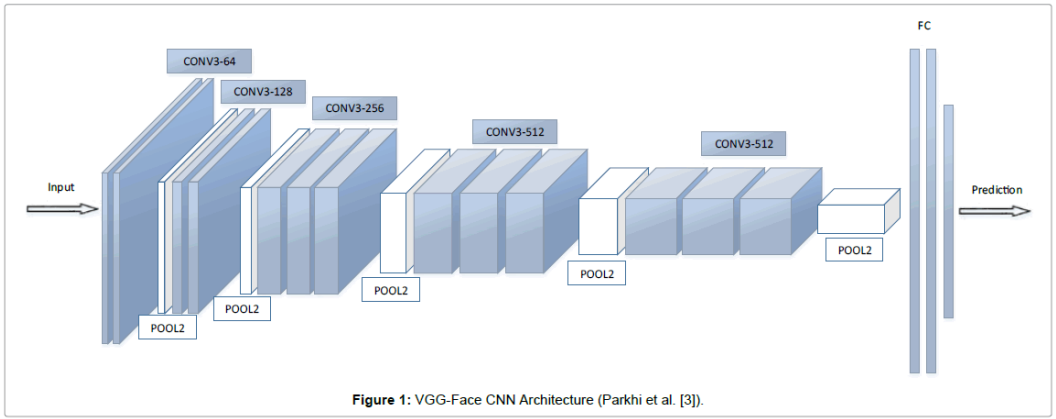
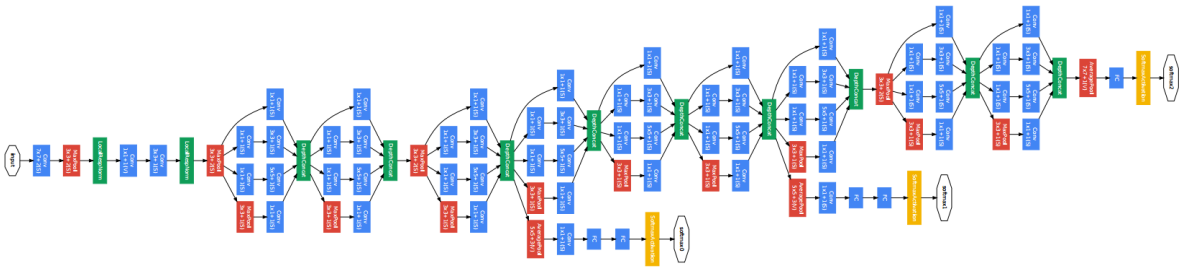


Figure 1: VGG-Face CNN Architecture (Parkhi et al. [3]).

Image classification (now)

- Deep CNNs
  - GoogleNet
  - ResNet



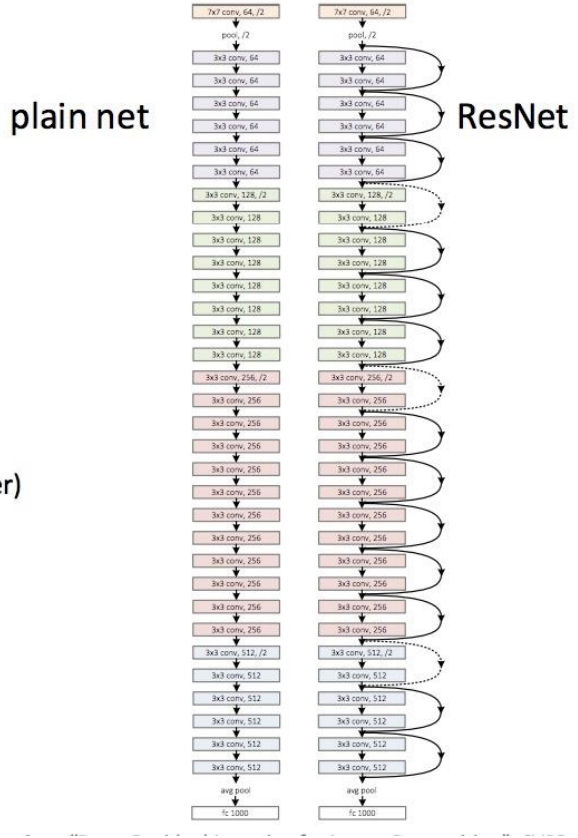
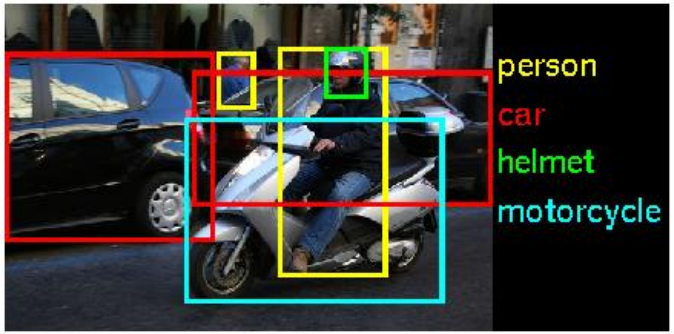


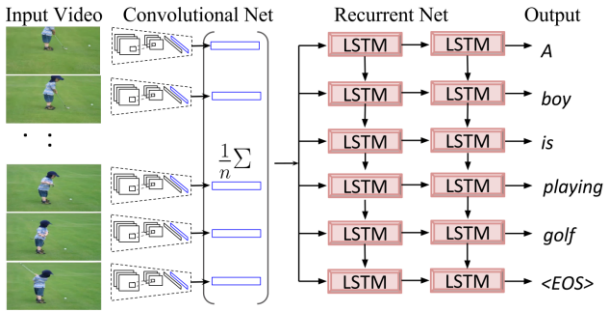
Image classification (now)

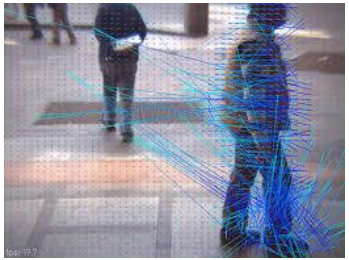
- Large labeled Datasets
  - ImageNet
  - COCO (Microsoft!)



Activity (now)

- Sliding window
- Optical flow
- LSTMs
- Temporal detection
- C3D





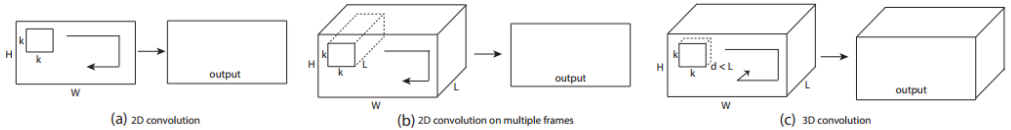


Figure 1. **2D and 3D convolution operations**. a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri,Learning Spatiotemporal Features with 3D Convolutional Networks, arXiv:1412.0767 [cs.CV]2015

Video datasets

- Sports 1M
- UCF 101
- ActivityNet
- cityscape
- Charades

Fig. 1. Comparison of actions in the Charades dataset and on YouTube: *Readi book*, *Opening a refrigerator*, *Drinking from a cup*. YouTube returns entertaining often atypical videos, while *Charades* contains typical everyday videos.

ACTIVITYNET

Image & Video Datasets

- Large-scale labeling is laborious
- Biased (usually augmentation is needed)
- Pixel-wise labeling is almost impossible
- Non-balanced classes occurances
- Abundant unlabeled online data !!

CG 4 CV

- Using Game Engines to generate data and use it to train NNs
- Works well only as augmentation (Domain Shift problem !)
- Still biased !
- PHAV: NIPS'16 : http://adas.cvc.uab.es/phav/

(a) RGB image (b) Semantic labels (c) Depth (d) Surface normals (e) Diffuse reflections

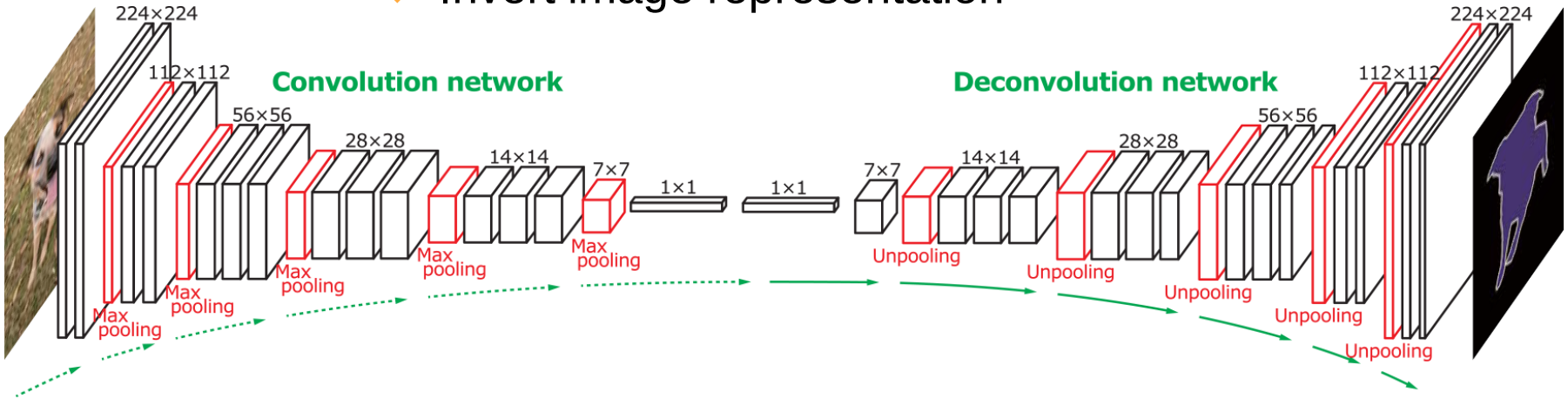
Figure 3: A rendered image sample together with corresponding pixel-level annotations.

outline

- Brief history (CNN , ImageNet, activity recognition)
- Brief history (DeconvNets , GAN, applications)
- My previous work on activity generation
- Recent GAN papers:
  - Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)
  - Adversarially Tuned Scene Generation (CVPR'17)
  - Generating Videos with Scene Dynamics (NIPS'16)
- What's next ?

Deconvolution Neural Networks

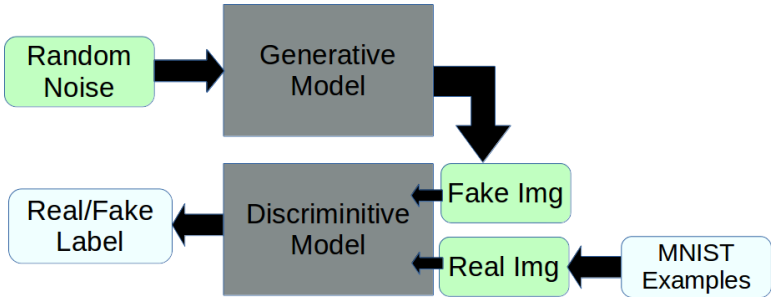
- Invert image representation



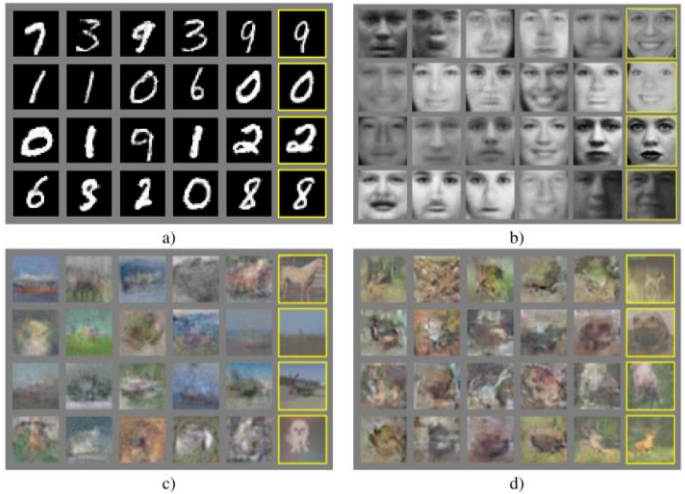
learning to generate chairs by CNN , Alexy Dovisky (CVPR15)

Deep Generative Adversarial Network

- Using Adversarial Loss have shown promising results (GAN)



$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \quad (1)$$



Generative Adversarial Nets , Ian J. Goodfellow, Jean Pouget-Abadie , NIPS'14

GAN applications

“GAN is the most important development of deep learning in C.V. after CNN “ *By Abdullah Hamdi*

Figure 22: Failure cases of type 3. Our photo-to-style transfer cannot produce geometry style transfer, like exaggerations of the eyes and the nose, due to the assumption in our work to preserve the content structure.

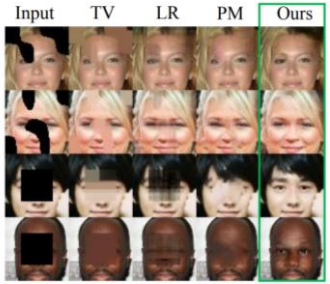


Figure 1. Semantic inpainting results by TV, LR, PM and our method. Holes are marked by black color.

- Super resolution
- Style transfer
- Complete lost parts

http://bamos.github.io/2016/08/09/deep-completion/

Semantic Image Inpainting with Deep Generative Models , Raymond A. Yeh∗ , Chen Chen∗ , Teck Yian Lim (arxive’17)

outline

- Brief history (CNN , ImageNet, activity recognition)
- Brief history (DeconvNets , GAN, applications)
- My previous work on activity generation
- Recent GAN papers:
  - Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)
  - Adversarially Tuned Scene Generation (CVPR'17)
  - Generating Videos with Scene Dynamics (NIPS'16)
- What's next ?

outline

- Brief history (CNN , ImageNet, activity recognition)
- Brief history (DeconvNets , GAN, applications)
- My previous work on activity generation
- Recent GAN papers:
  - Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)
  - Adversarially Tuned Scene Generation (CVPR'17)
  - Generating Videos with Scene Dynamics (NIPS'16)
- What's next ?

Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)

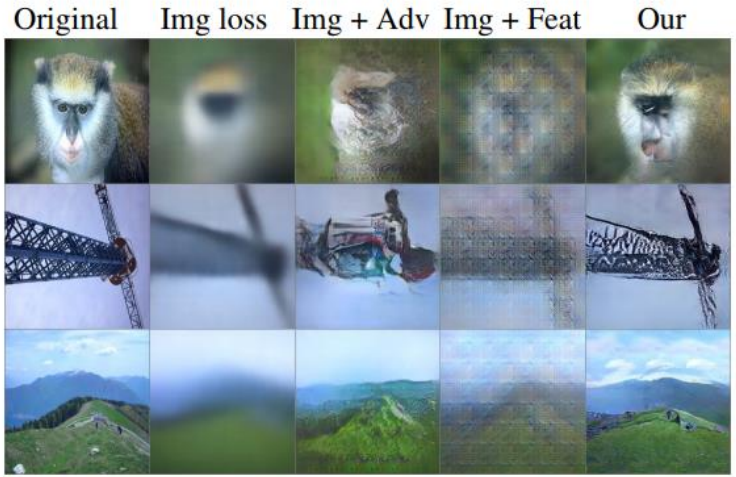
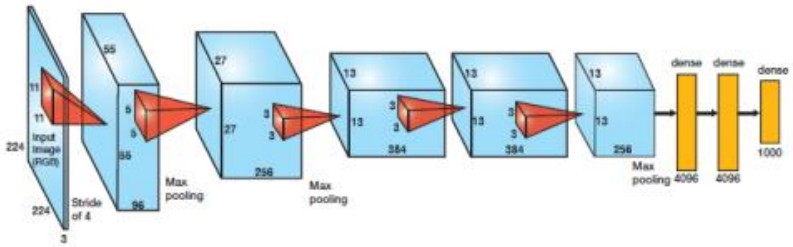


Figure 1: Reconstructions from AlexNet FC6 with different components of the loss.



Type	fc	fc	fc	reshape	uconv	conv	uconv	conv	uconv	conv	uconv	uconv	uconv
InSize	–	–	–	1	4	8	8	16	16	32	32	64	128
OutCh	4096	4096	4096	256	256	512	256	256	128	128	64	32	3
Kernel	–	–	–	–	4	3	4	3	4	3	4	4	4
Stride	–	–	–	–	↑2	1	↑2	1	↑2	1	↑2	↑2	↑2

Table 1: Generator architecture for inverting layer FC6 of AlexNet.

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16)

Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)

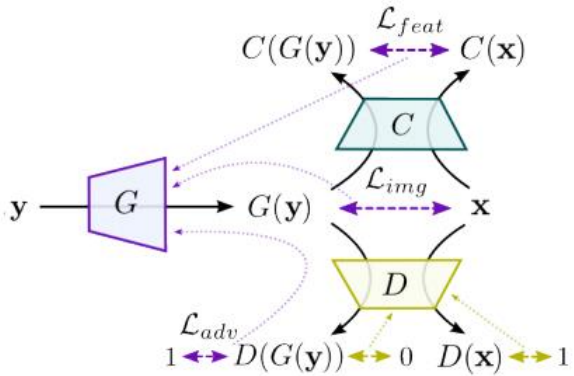


Figure 2: Schematic of our model. Black solid lines denote the forward pass. Dashed lines with arrows on both ends are the losses. Thin dashed lines denote the flow of gradients.

$$\mathcal{L} = \lambda_{feat}\,\mathcal{L}_{feat} + \lambda_{adv}\,\mathcal{L}_{adv} + \lambda_{img}\,\mathcal{L}_{img}.$$

$$\mathcal{L}_{feat} = \sum_i ||C(G_\theta(\mathbf{y}_i)) - C(\mathbf{x}_i)||_2^2. \tag{2}$$

$$\mathcal{L}_{discr} = -\sum_i \log(D_\varphi(\mathbf{x}_i)) + \log(1 - D_\varphi(G_\theta(\mathbf{y}_i))), \tag{3}$$

$$\mathcal{L}_{img} = \sum_i ||G_\theta(\mathbf{y}_i) - \mathbf{x}_i||_2^2. \tag{5}$$

$$\mathcal{L}_{adv} = -\sum_i \log D_\varphi(G_\theta(\mathbf{y}_i)). \tag{4}$$

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16)

Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS’16)

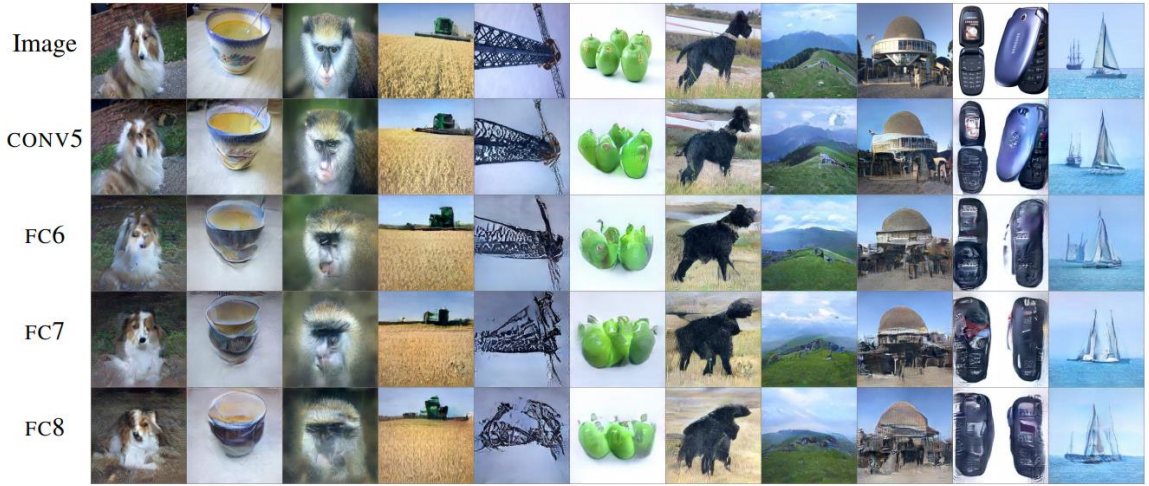


Figure 3: Representative reconstructions from higher layers of AlexNet. General characteristics of images are preserved very well. In some cases (simple objects, landscapes) reconstructions are nearly perfect even from FC8. In the leftmost column the network generates dog images from FC7 and FC8.

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS’16)

Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)

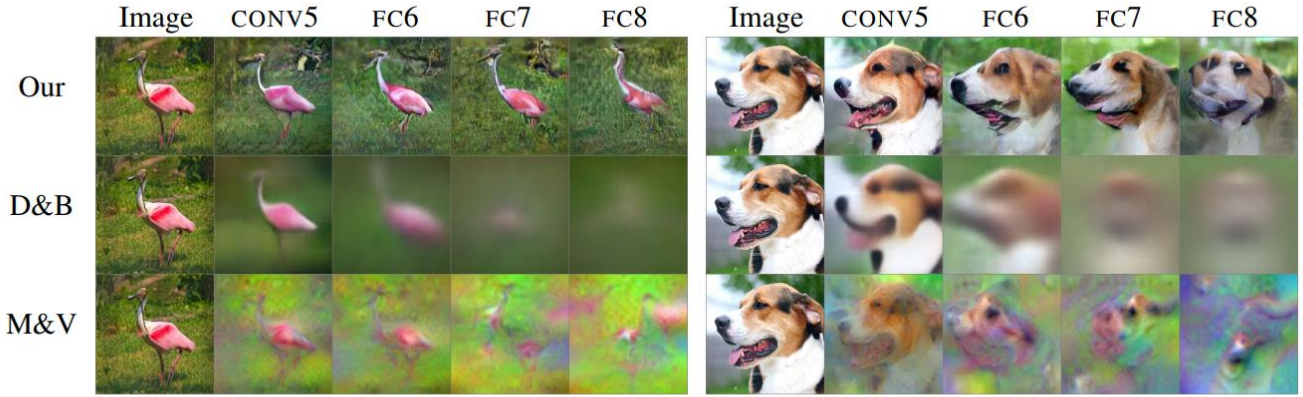


Figure 4: AlexNet inversion: comparison with Dosovitskiy and Brox [26] and Mahendran and Vedaldi [21] . Our results are significantly better, even our failure cases (second image).

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16)

Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)

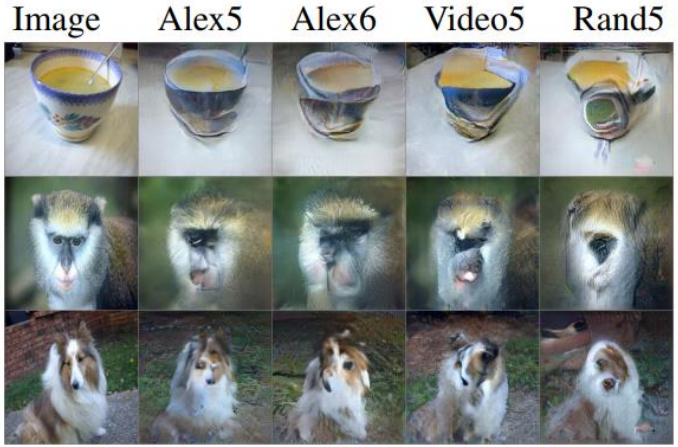


Figure 6: Reconstructions from FC6 with different comparators. The number indicates the layer from which features were taken.

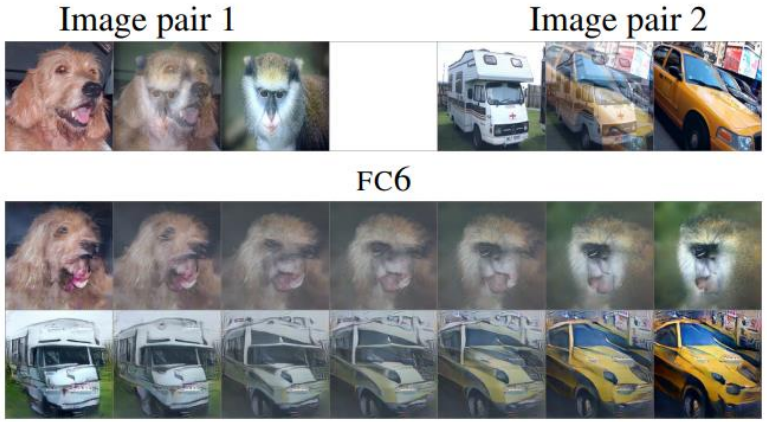


Figure 7: Interpolation between images by interpolating between their FC6 features.

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16)

Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)

- "For quantitative evaluation we compute the normalized Euclidean error ||a − b||/N. The normalization coefficient N is the average of Euclidean distances between all pairs of different samples from the test set."

- 0% -> exactly the inverse
- 100% -> similar to random image

	CONV5	FC6	FC7	FC8
M & V [21]	71/19	80/19	82/16	84/09
D & B [26]	35/–	51/–	56/–	58/–
Our image loss	–/–	46/79	–/–	–/–
AlexNet CONV5	43/37	55/48	61/45	63/29
VideoNet CONV5	–/–	51/57	–/–	–/–

Table 2: Normalized inversion error (in %) when reconstructing from different layers of AlexNet with different methods. First in each pair – error in the image space, second – in the feature space.

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16)

outline

- Brief history (CNN , ImageNet, activity recognition)
- Brief history (DeconvNets , GAN, applications)
- My previous work on activity generation
- Recent GAN papers:
  - Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)
  - Adversarially Tuned Scene Generation (CVPR'17)
  - Generating Videos with Scene Dynamics (NIPS'16)
- What's next ?

Adversarially Tuned Scene Generation (CVPR'17)

(e) A few samples from CityScapes data

(h) A few samples of $V_{cityscapes}$ sampled from the model after tuning

Adversarially Tuned Scene Generation, VSR Veeravasarapu , Constantin Rothkopf , Ramesh Visvanathan (CVPR'17)

Adversarially Tuned Scene Generation (CVPR'17)

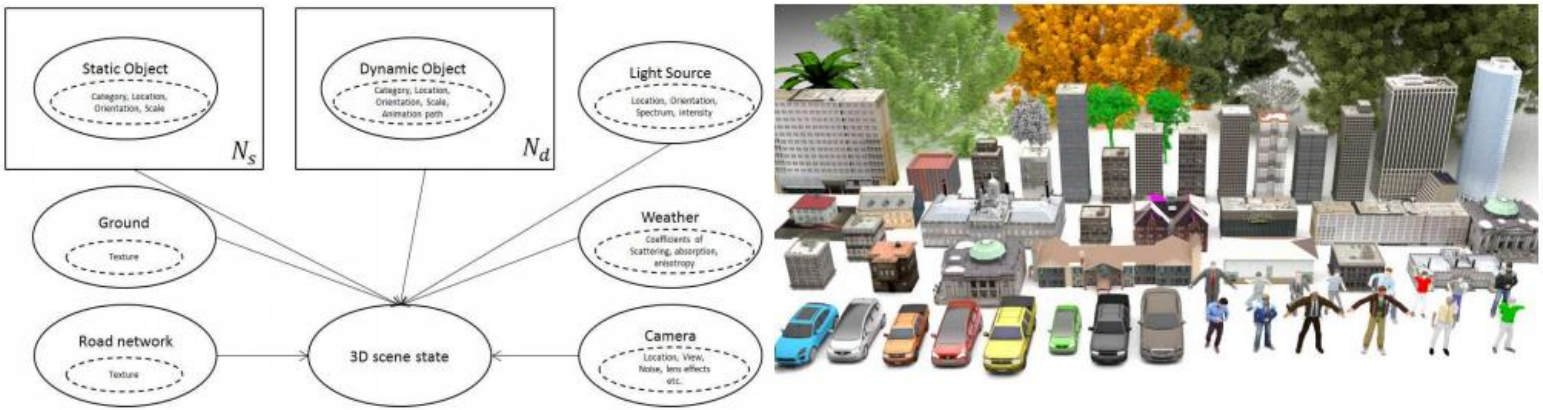


Figure 2: Graphical representation of the scene generative model and illustration of 3D CAD object models used in this work.

Adversarially Tuned Scene Generation, VSR Veeravasarapu , Constantin Rothkopf , Ramesh Visvanathan (CVPR'17)

Adversarially Tuned Scene Generation (CVPR'17)

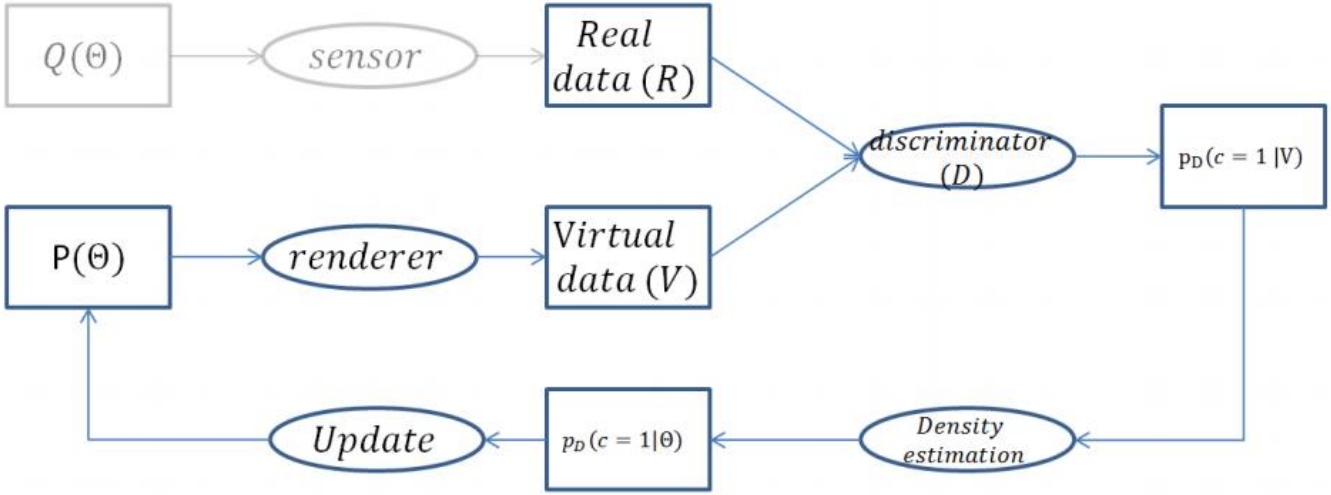


Figure 1: Flow chart of adversarial tuning

Adversarially Tuned Scene Generation, VSR Veeravasarapu , Constantin Rothkopf , Ramesh Visvanathan (CVPR'17)

(a) Histogram of V_{init} (b) A few samples of V_{init} sampled from the model before tuning (c) Pixel-proportions/class

(d) Histogram of CityScapes (e) A few samples from CityScapes data (f) Pixel-proportions/class

(g) Histogram of $V_{cityscapes}$ (h) A few samples of $V_{cityscapes}$ sampled from the model after tuning (i) Pixel-proportions/class

(j) Histogram of CamVid (k) A few samples from CamVid data (l) Pixel-proportions/class

(m) Histogram of V_{camvid} (n) A few samples of V_{camvid} sampled from the model after tuning (o) Pixel-proportions/class

Figure 4: Qualitative comparison of training sets, both simulated and real, and their statistics before and after tuning the generative model (Best viewed in color).

Adversarially Tuned Scene Generation (CVPR'17)

Table 1: Quantitative analysis of the performance of DeepLab models with different training-testing combinations. Notation: CS and CV refers to real CityScapes and CamVid datasets respectively, and prefix 'V' represents simulated sets.

Training set	*Validation*	*global*	*vehicle*	*pedestrian*	*building*	*vegetation*	*road*	*ground*	*sky*
Model Tuned to CityScapes data									
V_init	CS_val	49.86	48	53	63	51	47	34	53
V_cityscapes	CS_val	52.14 (+2.28)	56	47	65	57	53	31	56
CS_train	CS_val	67.71	59	57	73	64	69	64	88
V_cityscapes	CV_val	50.28 (+0.43)	51	50	55	48	49	49	50
CS_train	CV_val	54.42	47	43	55	69	46	51	70
Model Tuned to CamVid Data									
V_init	CV_val	46.42	53	38	54	35	43	39	63
V_camvid	CV_val	49.85 (+3.42)	57	34	63	37	48	44	66
CV_train	CV_val	67.42	77	34	65	54	98	45	99
V_camvid	CS_val	39.85 (-6.57)	35	41	44	44	32	40	43
CV_train	CS_val	54.28	46	43	55	69	46	51	70
Data augmentations									
V_init+10%CS	CS_val	67.42	60	66	52	67	74	72	81
V_cityscapes + 10%CS	CS_val	70.01 (+2.57)	68	60	59	68	77	69	89
V_init+10%CV	CV_val	68.85	51	61	71	67	65	77	90
V_camvid+10%CV	CV_val	70.57 (+1.71)	63	57	76	73	67	74	84

Adversarially Tuned Scene Generation, VSR Veeravasarapu , Constantin Rothkopf , Ramesh Visvanathan (CVPR'17)

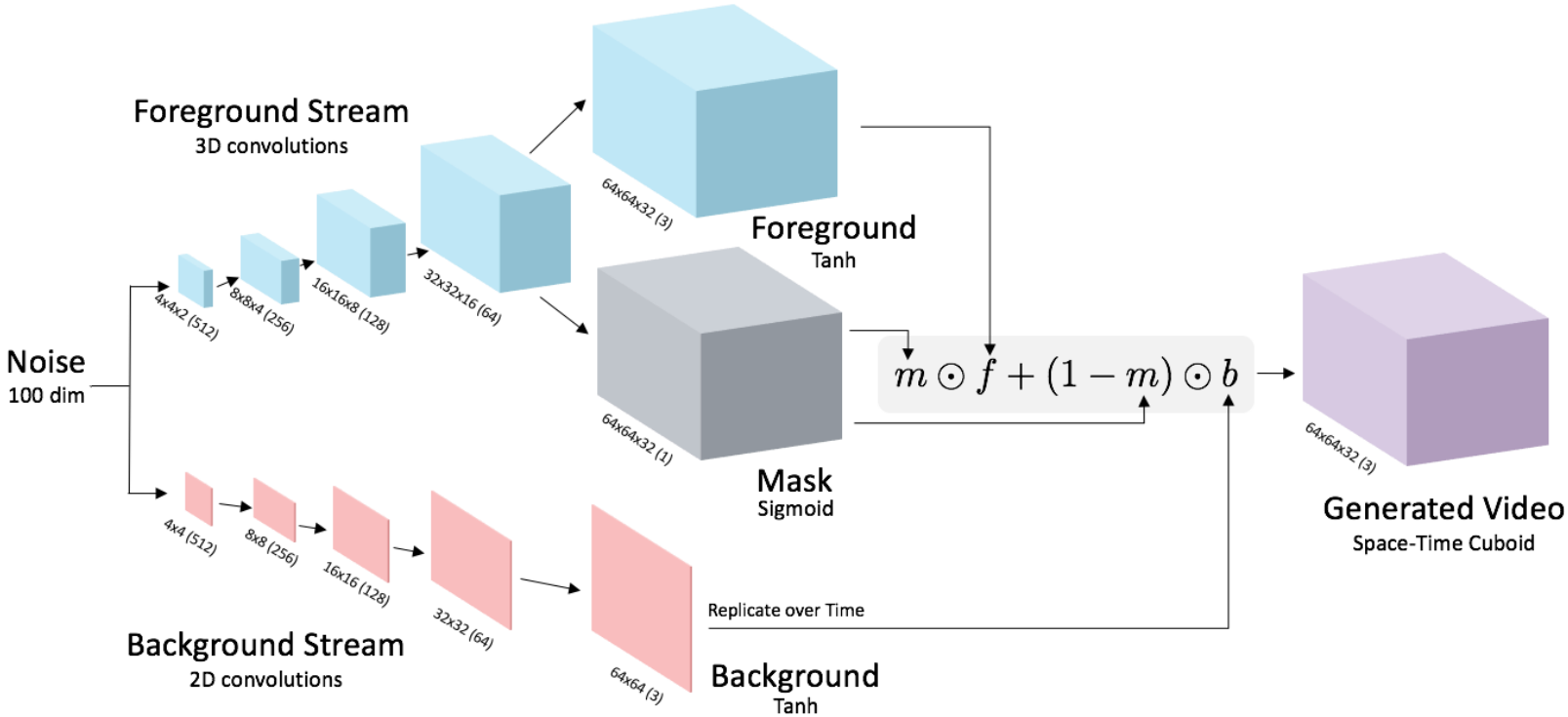
outline

- Brief history (CNN , ImageNet, activity recognition)
- Brief history (DeconvNets , GAN, applications)
- My previous work on activity generation
- Recent GAN papers:
  - Generating Images with Perceptual Similarity Metrics based on Deep Networks(NIPS'16)
  - Adversarially Tuned Scene Generation (CVPR'17)
  - Generating Videos with Scene Dynamics (NIPS'16)
- What's next ?

Generating Videos with Scene Dynamics (NIPS'16)

Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

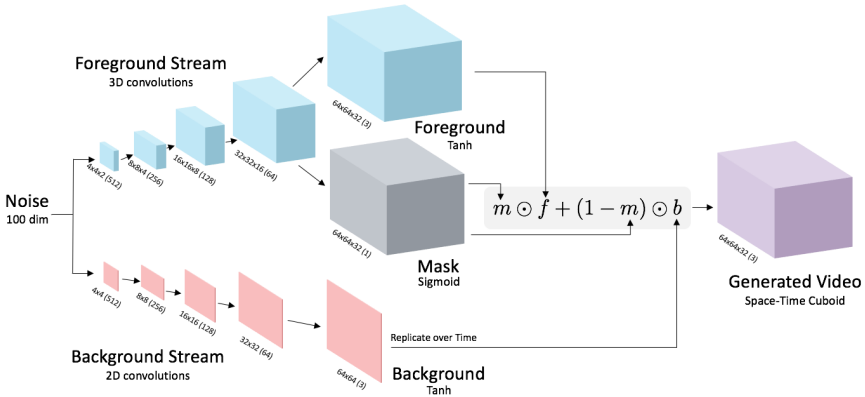
Generating Videos with Scene Dynamics (NIPS'16)



Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

Generating Videos with Scene Dynamics (NIPS'16)

$$\min_{w_G} \max_{w_D} \mathbb{E}_{x \sim p_x(x)} [\log D(x; w_D)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z; w_G); w_D))] \quad (1)$$



$$G_2(z) = m(z) \odot f(z) + (1 - m(z)) \odot b(z).$$

Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

Generating Videos with Scene Dynamics (NIPS'16)

- modeling scene dynamics is almost impossible by hand for prediction
- separate scene into background and object with two different pipelines .. in activity recognition many times the background is stationary

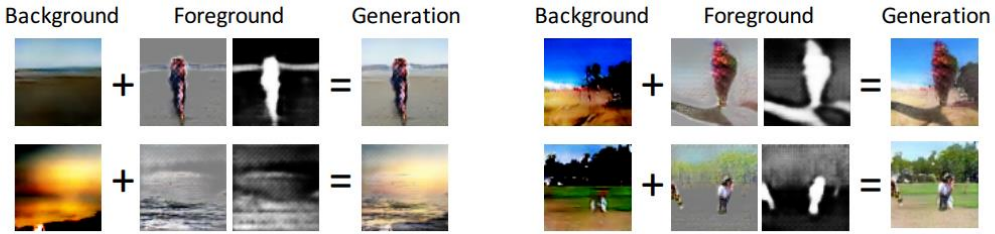


Figure 3: **Streams:** We visualize the background, foreground, and masks for beaches (left) and golf (right). The network generally learns to disentangle the foreground from the background.

Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

Generating Videos with Scene Dynamics (NIPS'16)

- Video generation

"Which video is more realistic?"	Percentage of Trials				
	Golf	Beach	Train	Baby	Mean
Random Preference	50	50	50	50	50
Prefer VGAN Two Stream over Autoencoder	88	83	87	71	82
Prefer VGAN One Stream over Autoencoder	85	88	85	73	82
Prefer VGAN Two Stream over VGAN One Stream	55	58	47	52	53
Prefer VGAN Two Stream over Real	21	23	23	6	18
Prefer VGAN One Stream over Real	17	21	19	8	16
Prefer Autoencoder over Real	4	2	4	2	3

Table 1: **Video Generation Preferences:** We show two videos to workers on Amazon Mechanical Turk, and ask them to choose which video is more realistic. The table shows the percentage of times that workers prefer one generations from one model over another. In all cases, workers tend to prefer video generative adversarial networks over an autoencoder. In most cases, workers show a slight preference for the two-stream model.

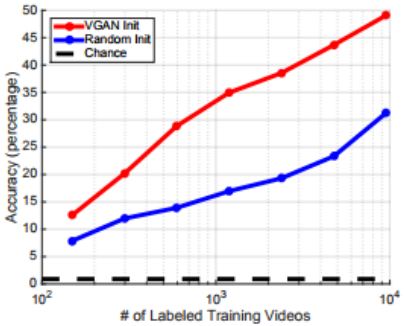
Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

Generating Videos with Scene Dynamics (NIPS'16)

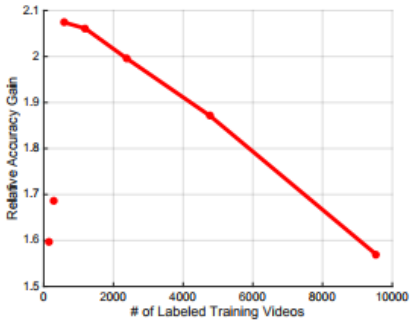
- Activity recognition

Method	Accuracy
Chance	0.9%
STIP Features [36]	43.9%
Temporal Coherence [10]	45.4%
Shuffle and Learn [25]	50.2%
VGAN + Random Init	36.7%
VGAN + Logistic Reg	49.3%
VGAN + Fine Tune	**52.1%**
ImageNet Supervision [47]	91.4%

(a) Accuracy with Unsupervised Methods



(b) Performance vs # Data

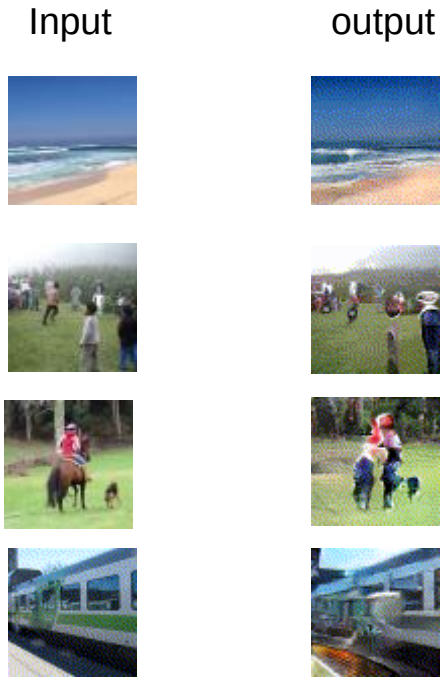


(c) Relative Gain vs # Data

Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

Generating Videos with Scene Dynamics (NIPS'16)

- Future prediction



Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

Thank you!

^_^

Useful links

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri,Learning Spatiotemporal Features with 3D Convolutional Networks, arXiv:1412.0767 [cs.CV]2015

Generative Adversarial Nets , Ian J. Goodfellow, Jean Pouget-Abadie , NIPS'14 https://arxiv.org/pdf/1406.2661.pdf

learning to generate chairs by CNN , Alexy Dovisky (CVPR15)

https://www.robots.ox.ac.uk/~vgg/rg/papers/Dosovitskiy_Learning_to_Generate_2015_CVPR_paper.pdf

Semantic Image Inpainting with Deep Generative Models , Raymond A. Yeh∗ , Chen Chen∗ , Teck Yian Lim (arxive'17) https://arxiv.org/pdf/1607.07539.pdf

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16) . http://papers.nips.cc/paper/6158-generating-images-with-perceptual-similarity-metrics-based-on-deep-networks.pdf

Useful links (2)

Generating Images with Perceptual Similarity Metrics based on Deep Networks, Alexey Dosovitskiy and Thomas Brox(NIPS'16) . http://papers.nips.cc/paper/6158-generating-images-with-perceptual-similarity-metrics-based-on-deep-networks.pdf

Generating Videos with Scene Dynamics, Carl Vondrick, Hamed Pirsiavash (NIPS'16)

http://carlvondrick.com/tinyvideo/paper.pdf

Adversarially Tuned Scene Generation, VSR Veeravasarapu , Constantin Rothkopf , Ramesh Visvanathan (CVPR'17)
http://openaccess.thecvf.com/content_cvpr_2017/papers/Veeravasarapu_Adversarially_Tuned_Scene_CVPR_2017_paper.pdf