What have we learned from deep representations for action recognition?

Christoph Feichtenhofer, Axel Pinz, Richard P. Wildes, Andrew Zisserman

CVPR 2018

Sofia Broomé, CV/DL Reading group 4 Dec 2018

In summary

- First to visualize **hierarchical features** in deep **motion** network
- **Activity maximization**
- Spatial and temporal **regularization**
- Both specific and generic **representations**

Interpretability timeline

Erhan 2009
Introduces activity maximization

Zeiler 2013
Uses deconvnets (their invention from 2010) to project the feature activations back to pixel space

Szegedy 2013
1. Questions the semantic interp. of single units ”grandmother cells”, claims it is rather a distributed code
2. First to look at (and coins) adversarial examples

Agrawal 2014
Agrees with the distributed code argument and presents more experiments to support this

Simonyan 2014
1. Applies act. max. on a supervised convnet model
2. Computes saliency maps using backprop
3. Shows that such gradient-based vis. methods generalize deconvolution reconstructions

Interpretability timeline

Zhou 2015
Introduces class activation mapping (CAM)

Mahendran 2016
Another questioning of the grandmother cells. This time supported by the notion of the information bottleneck.

Selvaraju 2017
Grad-CAM
More general solution than CAM

Selvaraju 2018
Grad-CAM++
More robust to when there are multiple instances to classify
Applied to video

Feichtenhofer 2018
Focuses on spatiotemporal visualizations (for 2-stream models), the approach is without given input (activity maximization)

Adds regularization

+ Have moving video in their pdf paper 👍
- Ignores no-GMC papers

Mahendran, A., & Vedaldi, A. (2016). Salient deconvolutional networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9910 LNCS*, 120–135.

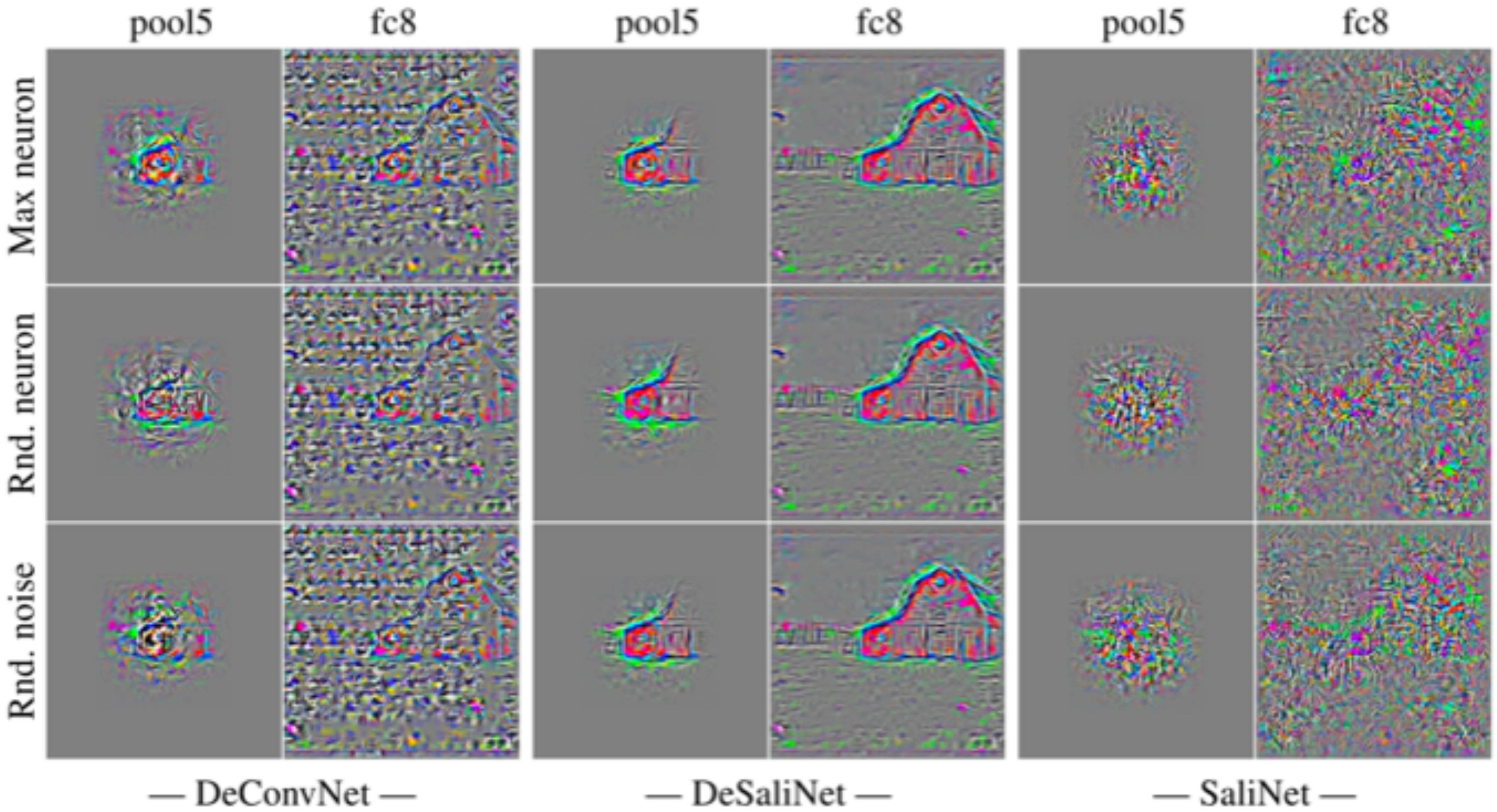
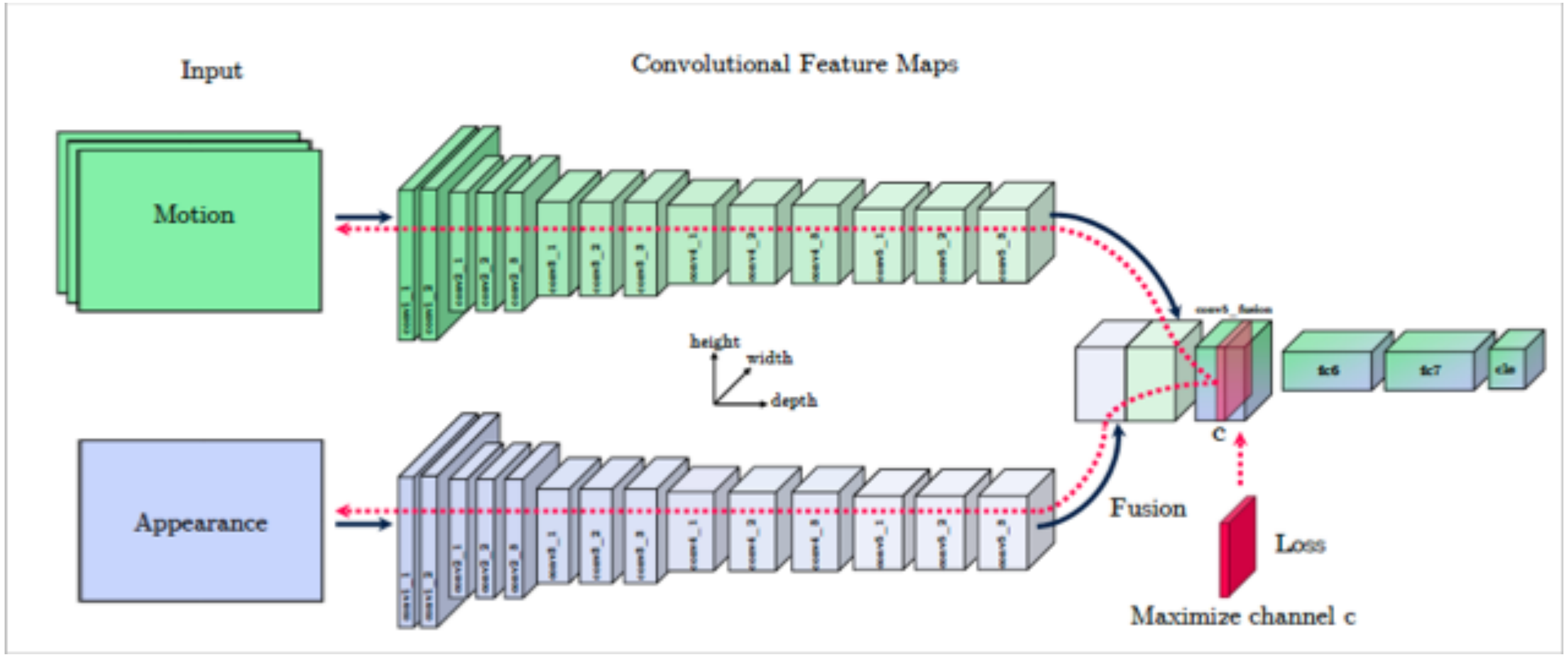


Fig. 4: *Lack of neuron selectivity.* The bottleneck information $\mathbf{r}$ is fixed to the one computed during the forward pass $\phi(\mathbf{x})$ through AlexNet and the output of $\phi^{\dagger}(\mathbf{e}, \mathbf{r})$ is computed by choosing $\mathbf{e}$ as: the most active neuron (top row), a second neuron at random (middle), or as a positive random mixture of all neurons (bottom row). Results barely differ, particularly for the deeper layers. See figure 1 for the original house input image $\mathbf{x}$. Best viewed on screen.

Approach – Activity maximixation



$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmax}} \frac{1}{\rho_l^2 \hat{\mathbf{a}}_{l,c}} \langle \mathbf{a}_l(\mathbf{x}), e_c \rangle - \lambda_r \mathcal{R}_r(\mathbf{x}) \qquad (1)$$

Approach – Regularizing local energy

$$\mathcal{R}_B(\mathbf{x}) = \begin{cases} N_B(\mathbf{x}) & \forall i, j, k : \sqrt{\sum_d \mathbf{x}(i,j,k,d)^2} \leq B \\ +\infty, & \text{otherwise.} \end{cases} \tag{2}$$

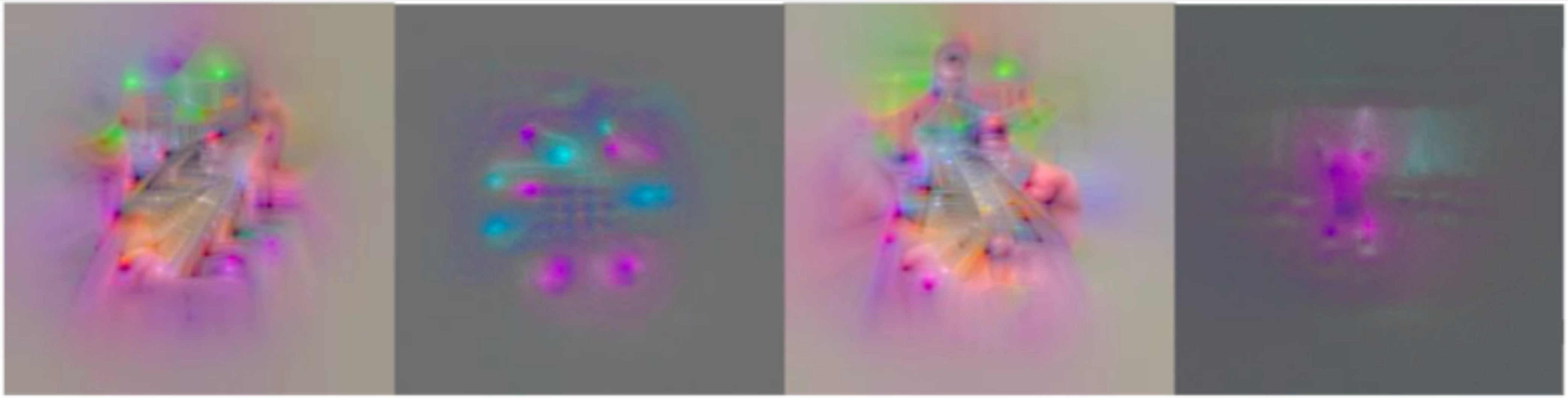
$$N_B(\mathbf{x}) \;=\; \sum_{i,j} \left(\sum_d \mathbf{x}(i,j,k,d)^2\right)^{\frac{\alpha}{2}}$$

Approach – Regularizing local frequency

$$\mathcal{R}_{TV}(\mathbf{x};\kappa,\chi) = \sum_{ijkd}\left[\kappa\left((\nabla_x \mathbf{x})^2 + (\nabla_y \mathbf{x})^2\right) + \chi(\nabla_t \mathbf{x})^2\right], \quad (3)$$

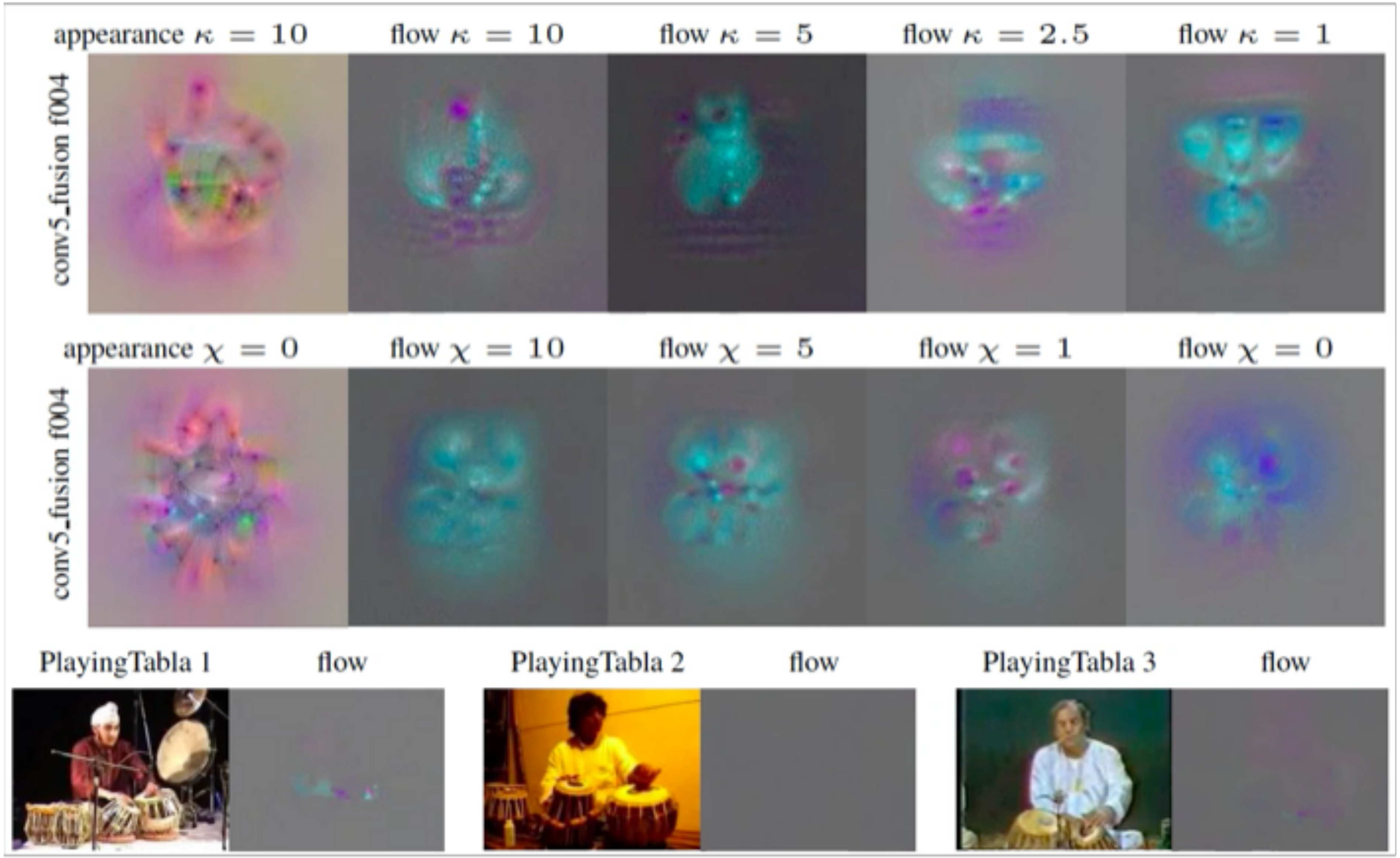
By varying χ and **κ** we can have 3 different cases:

- A purely spatial regularizer (**κ** > 0; $\chi = 0$)
- An isotropic spatiotemporal regularizer (**κ** $= \chi$; $\chi > 0$)
- An anisotropic spatiotemporal regularizer (**κ** $\neq \chi$; **κ** $, \chi > 0$)

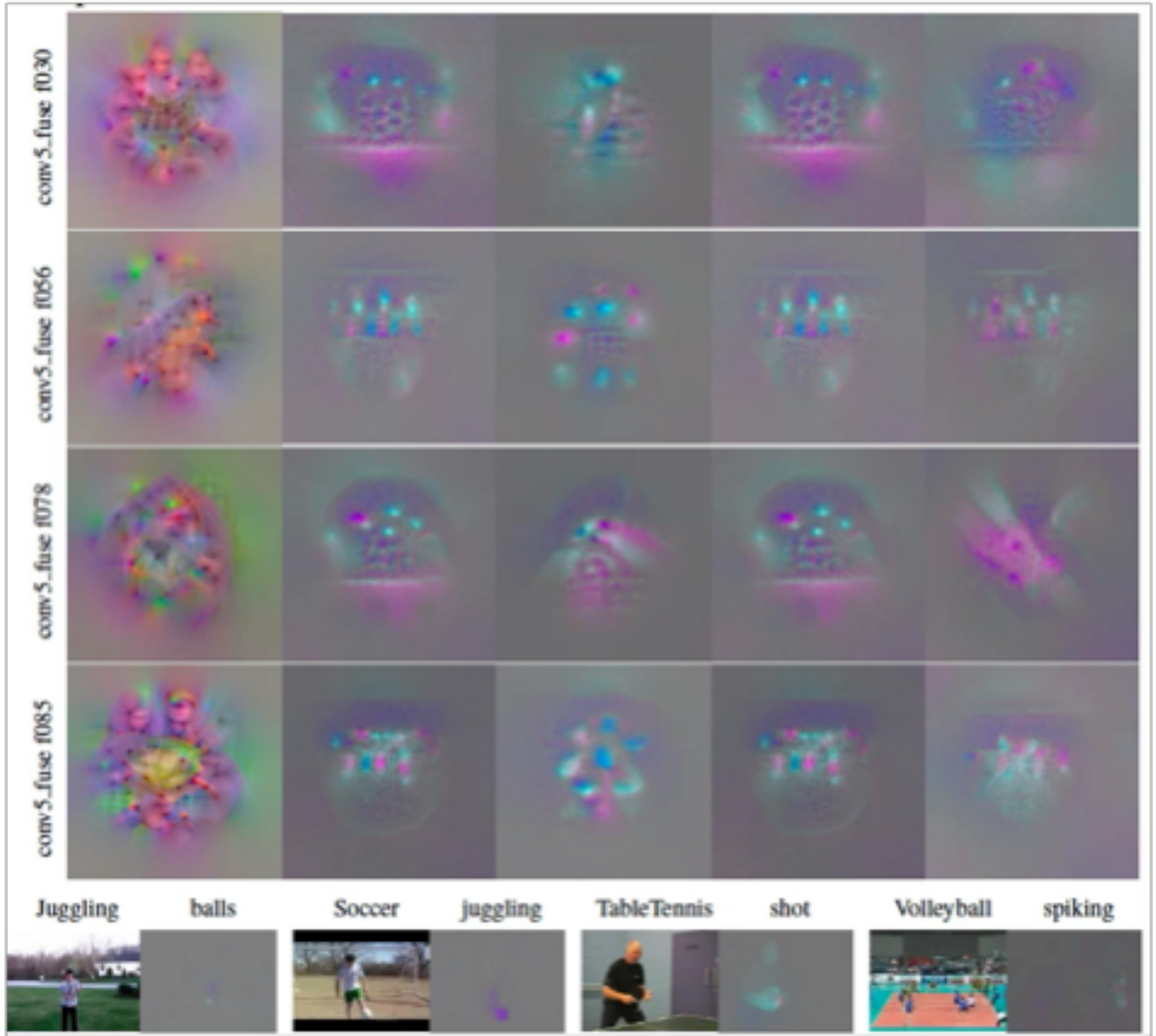


(a) appearance slow (b) motion slow (c) appearance fast (d) motion fast

Experiments – "Class-specific" units



Experiments – General units



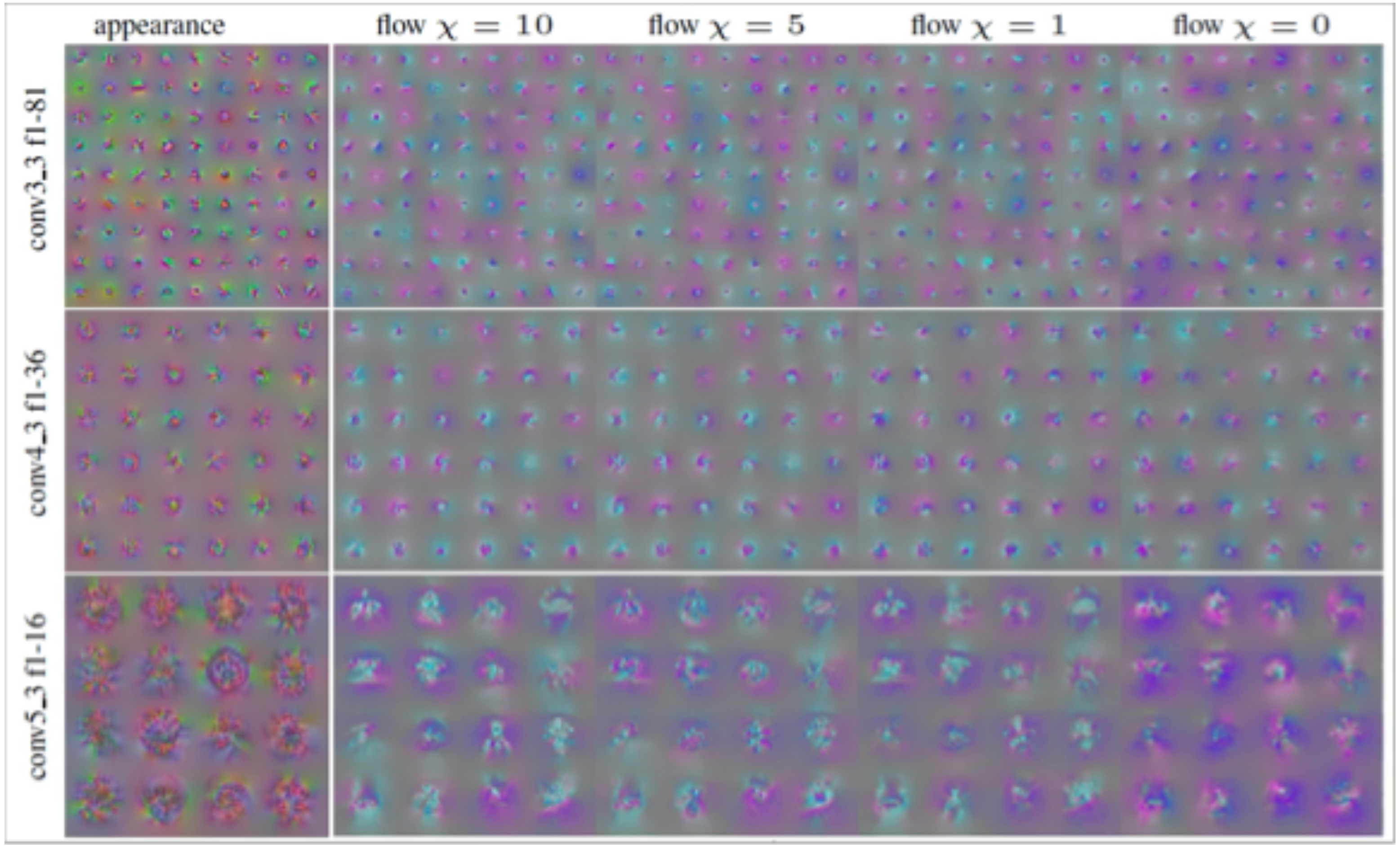
Experiments – Progressive feature abstraction with depth

Early layers:

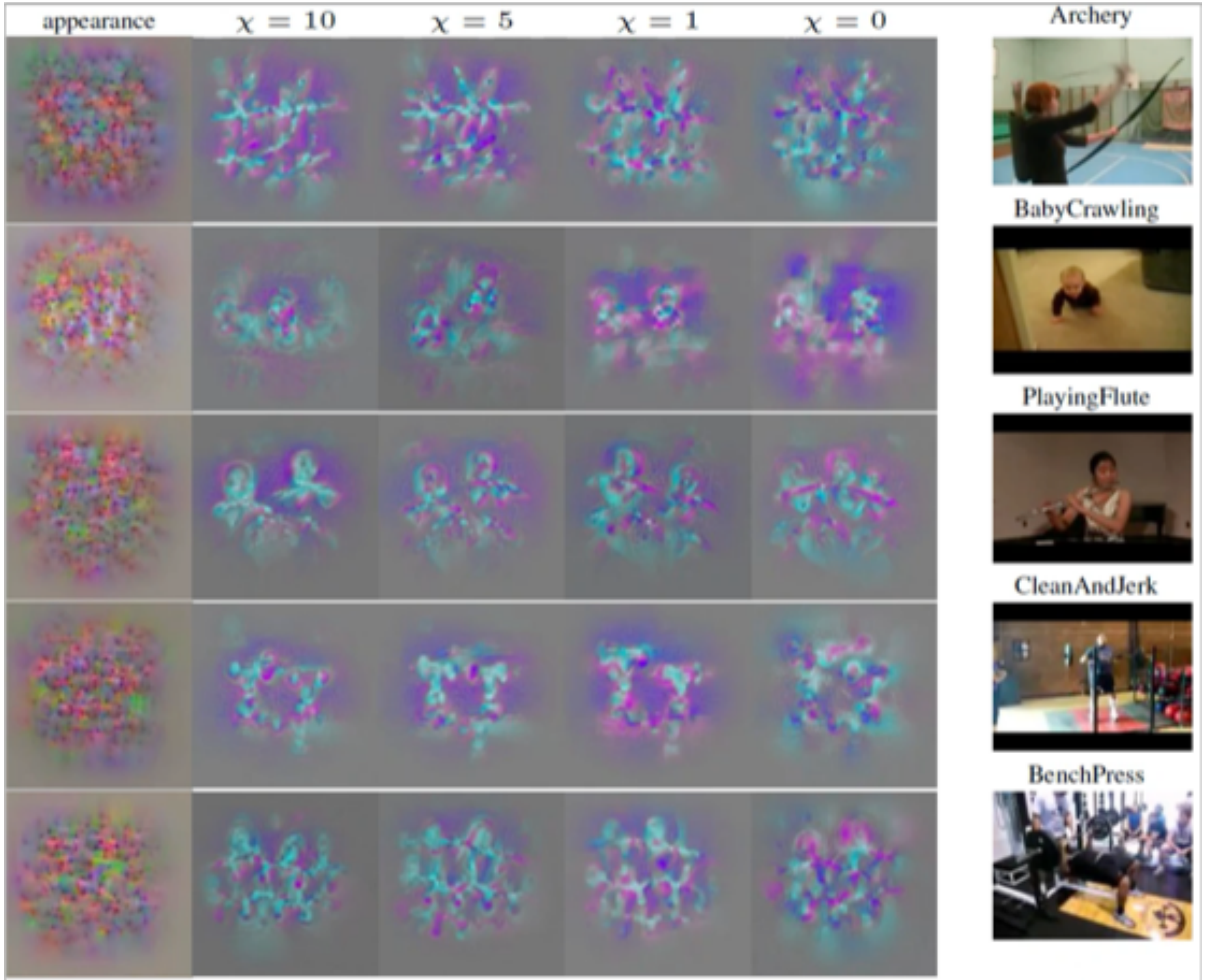
- Spatial patterns preserved regardless of Chi
- Speed invariance

Fusion layers:

Same results as earlier, matching to classes by inspection



Experiments – Progressive feature abstraction with depth



Global layers:

- Matching by inspection, varying Chi

Class output layer:

- Know what the units should correspond to
- Motion is striking, appearance not as specific (faces, barbells)

Claims to discuss

- "Our visual explanations provide qualitative support for the benefits of separating into two pathways when processing spatiotemporal information"
- "At conv5_fusion we see the emergence of both class specific and class agnostic units"

More?