Convolutional Pose Machines

What are CPMs?

Convolutional Pose Machines (CPMs) are deep learning architectures introduced in 2016 for **human pose estimation**, detecting and locating body parts in images. They use a sequence of CNNs to iteratively refine joint localization.

What are CPMs?

CPMs are crucial in computer vision applications like human-computer interaction, and are a founding block to other important pose estimation algorithms like **Multi-Person using Part Affinity Fields**

Too simplistic :(

The classical method for estimating body poses, called the pictorial structures model, uses a tree-like structure to understand how body parts connect and move together. This approach works well when all body parts are visible but struggles with mistakes like double-counting parts because it can't capture all correlations.

Comparison with older methods

Too difficult :(

Other methods, like hierarchical models, use larger body parts to help locate smaller ones, while non-tree models add extra connections to handle symmetry and occlusion. These methods often rely on approximate calculations for efficiency, while newer methods learn complex interactions directly through sequential predictions.

Comparison with older methods

Who cares about explicit structure!

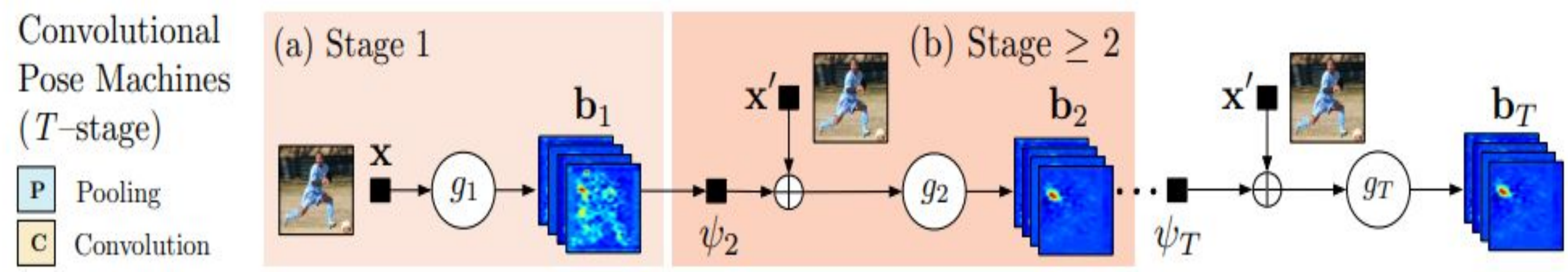
Who cares about explicit structure!

Convolutional Pose Machines (CPMs) changed the game for pose estimation by **using deep learning** to understand body part relationships **without needing manually designed rules.** Unlike older methods, which either directly guessed body part locations or needed complicated setups, CPMs use a series of neural networks to **refine their guesses step by step.** This allows CPMs to **understand complex spatial details** and avoid common problems.----------------------> better **accuracy** and **efficiency** in detecting body poses compared to previous techniques.

What Are Pose machines???

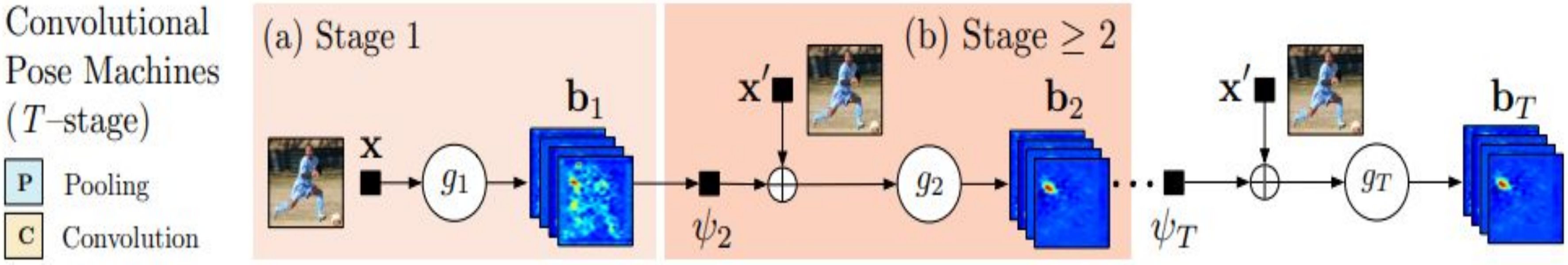
(*NORMAL*) POSE MACHINES

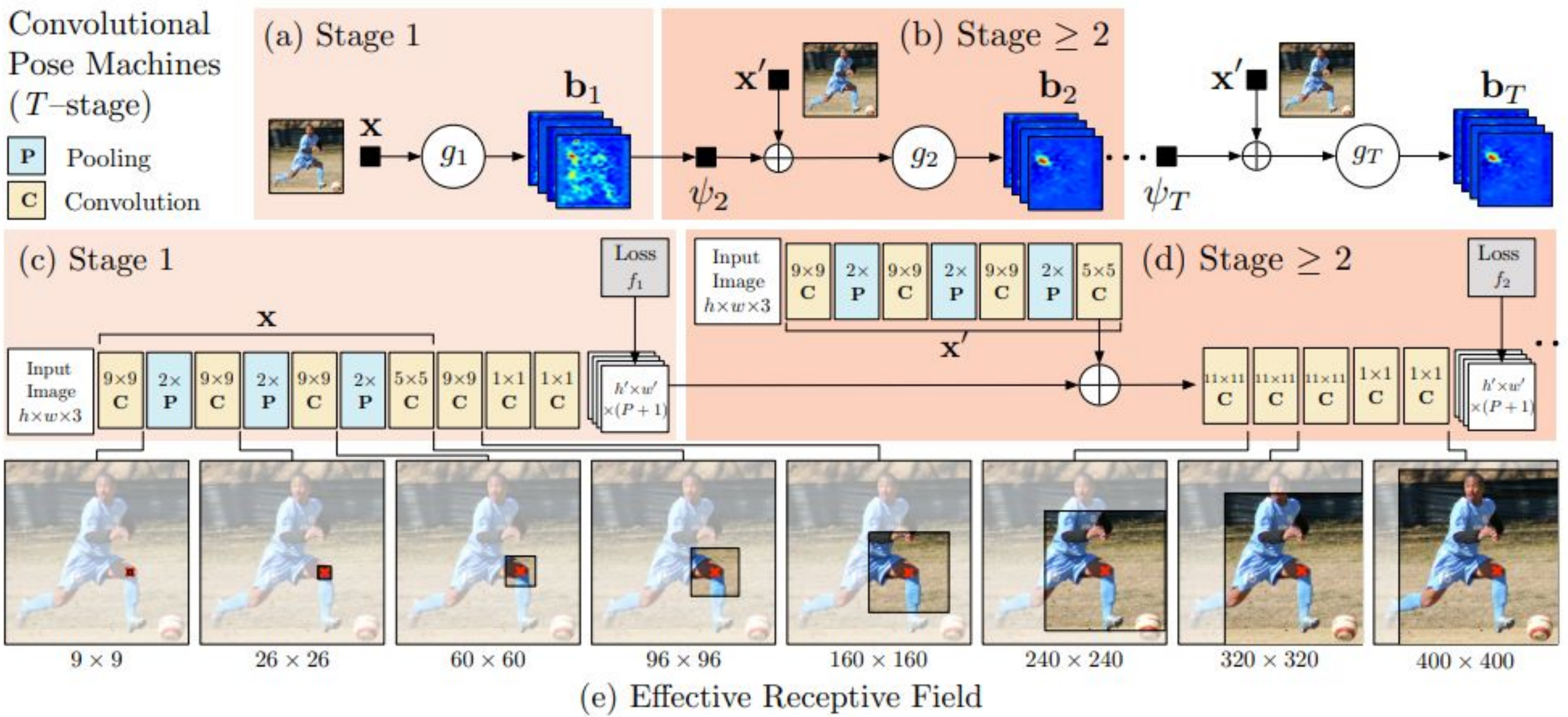
A pose machine predicts the locations of body parts in an image using a sequence of classifiers. Each classifier, at different stages, makes "beliefs" or guesses about where each body part is based on image features and information from the previous stage. The first stage uses just the image features to make initial guesses, while later stages refine these guesses using both the image features and the previous stage's beliefs.



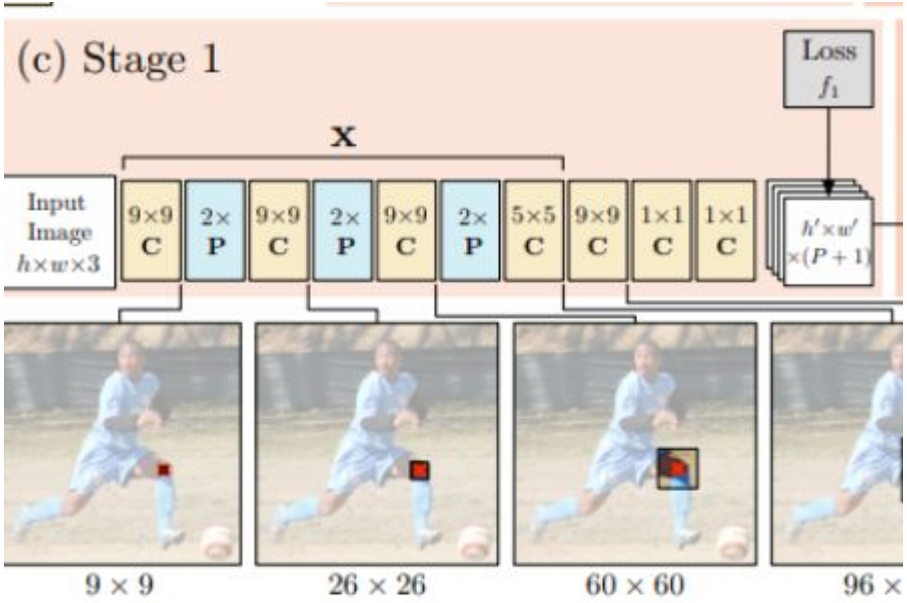
CONVOLUTIONAL POSE MACHINES

Initially, pose machines used boosted **random forests** for predictions and fixed hand-crafted features to understand spatial context. Convolutional Pose Machines (CPMs) improved on this by using convolutional neural networks (CNNs) instead. This allows CPMs to learn features directly from the data and refine pose estimates more accurately through multiple stages of CNNs, resulting in better performance and precision.



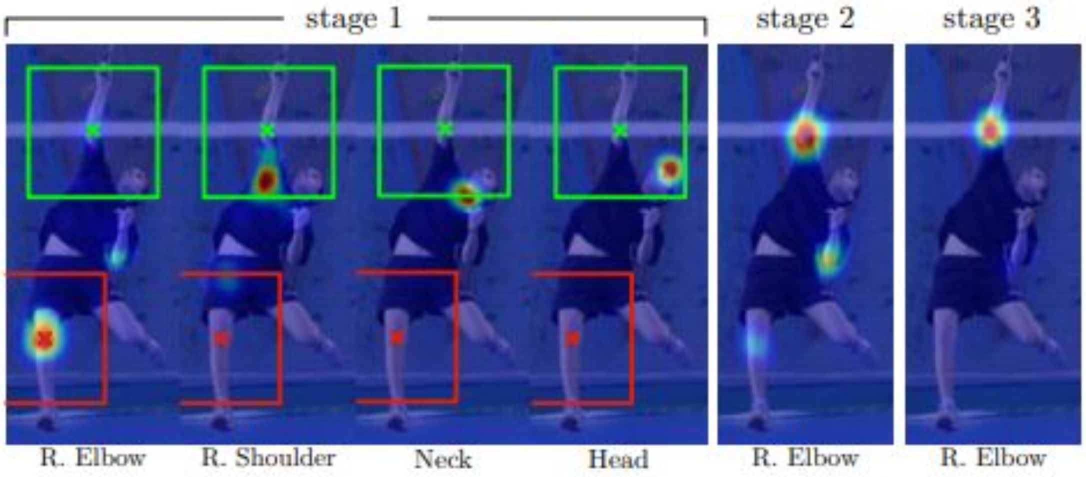

Convolutional Pose Machines (T–stage)
P Pooling
C Convolution
(a) Stage 1
$\mathbf{x}$
g_1
$\mathbf{b}_1$
(b) Stage ≥ 2
$\mathbf{x}'$
ψ_2
g_2
$\mathbf{b}_2$
ψ_T
g_T
$\mathbf{b}_T$
(c) Stage 1
Loss f_1
Input Image $h \times w \times 3$
9×9 C
2× P
9×9 C
2× P
9×9 C
2× P
5×5 C
9×9 C
1×1 C
1×1 C
$h' \times w' \times (P+1)$
(d) Stage ≥ 2
Loss f_2
Input Image $h \times w \times 3$
9×9 C
2× P
9×9 C
2× P
9×9 C
2× P
5×5 C
$\mathbf{x}'$
11×11 C
11×11 C
11×11 C
1×1 C
1×1 C
$h' \times w' \times (P+1)$
9 × 9
26 × 26
60 × 60
96 × 96
160 × 160
240 × 240
320 × 320
400 × 400
(e) Effective Receptive Field

At first small receptive field (*local*)



Convolutional Pose Machines (CPMs) replace the prediction and image feature computation modules of traditional pose machines with a **deep convolutional architecture.** This allows CPMs to learn both image and contextual features **directly from data** and enables **end-to-end** joint training of all stages. The **first stage** of a CPM predicts part locations **using local image evidence** from a small patch around each pixel, employing a network with five convolutional layers followed by two 1x1 convolutional layers.

LATER BIGGER RECEPTIVE FIELD (*GLOBAL*)



In later stages, CPMs improve part location predictions **using information from earlier stages**. For example, if the first stage accurately predicts **the shoulder**, this helps locate **harder parts like the elbow.** The second stage uses a larger receptive field to **learn complex relationships** between parts and refine predictions. By combining local evidence and spatial context through deep layers, CPMs achieve high accuracy in pose estimation. ... **THE LARGER THE RECEPTIVE BECOMES THE BETTER!**

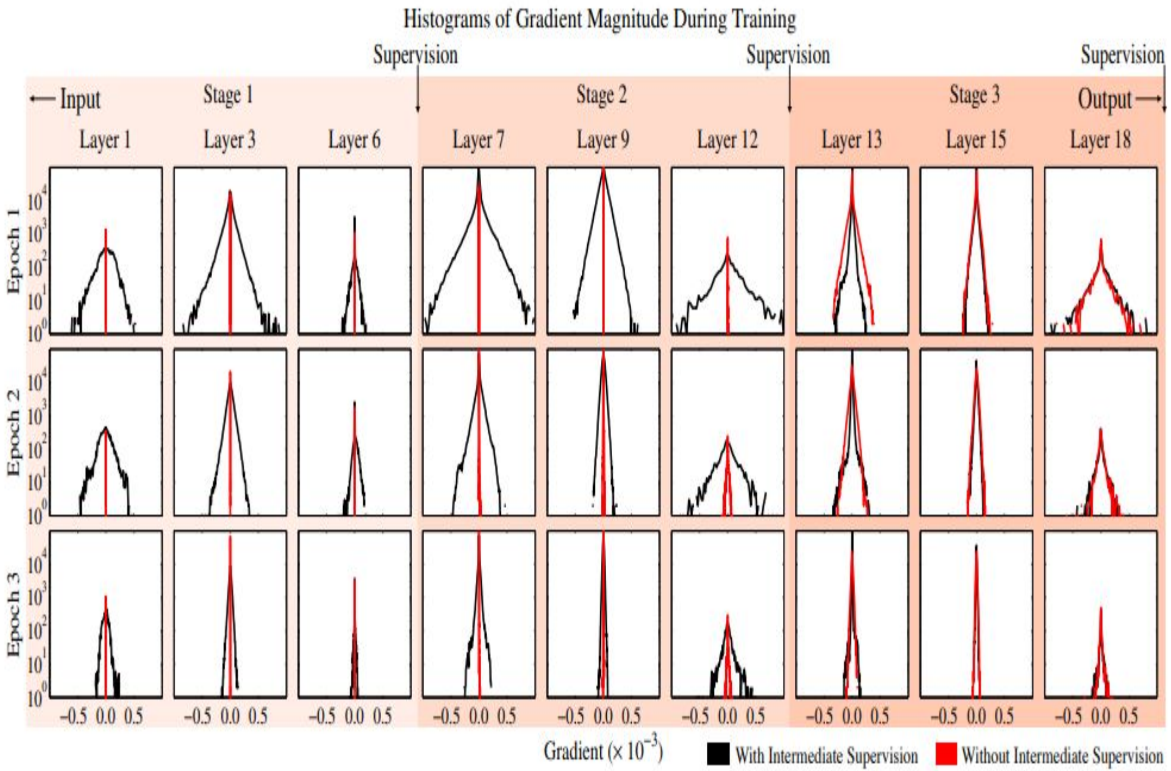
One more great thing !

INTERMEDIATE SUPERVISION!!! NO GRADIENT GOING POOF!

Convolutional Pose Machines (CPMs) are particularly effective for learning and addressing the vanishing gradient problem through a technique called **intermediate supervision**. In CPMs, the loss function is applied not just at the final output but at each stage of the network. This means that during training, gradients are computed and used **at multiple stages,** which prevents the gradients from becoming too small as they pass through the network's layers—**a common issue** in **deep networks** known as the vanishing gradient problem.

Intermediate supervision

By applying supervision at each stage, CPMs ensure that learning occurs throughout the entire network. This approach helps maintain a strong gradient signal, which is crucial for training deep networks effectively. Studies show that networks with intermediate supervision have a wider range of gradient magnitudes across layers, which supports better learning early in training and leads to effective model convergence over time.



THANK YOU! <3