Deep High-Resolution Representation Learning for Human Pose Estimation

Ke Sun[1] , Bin Xiao[2], Dong Liu[1], Jingdong Wang[2],

[1] University of Science and Technology of China [2] Microsoft Research Asia

(CVPR 2019)

HyunJae Bae

jason0425@g.skku.edu

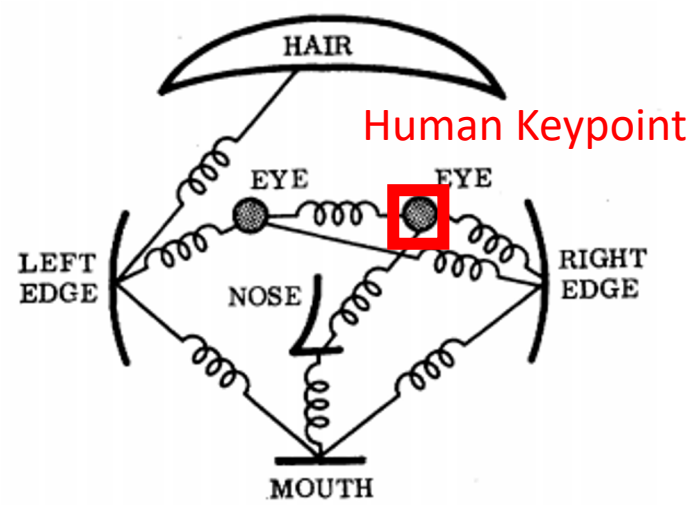
SKKU Information & Intelligence System Lab

2021/08/11

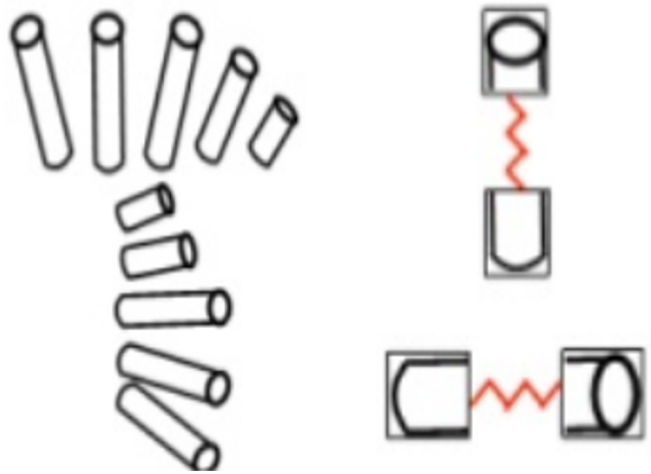
Background

- Human Pose Estimation

  To localize human anatomical keypints or parts(e.g. elbow, writs, etc.)



Fischler & Elschlager 1973



Yang & Ramanan 2011

M. Eichner & V. Ferrari 2009

Pictorial Structure

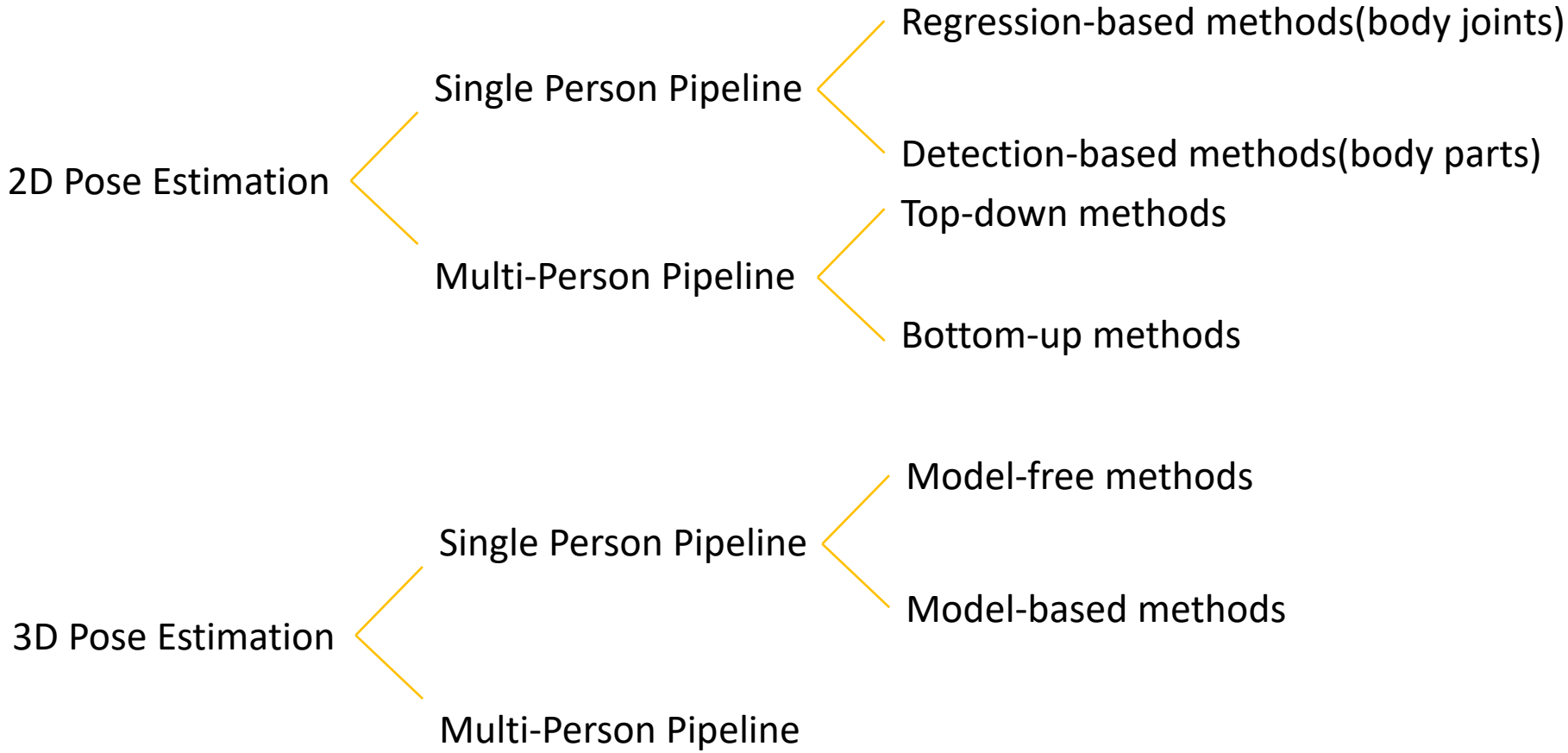
- Pairwise Springs
- Spring Tension(Cost) ∝ Distance

Mixture of "mini-parts"

- Mixture of part i

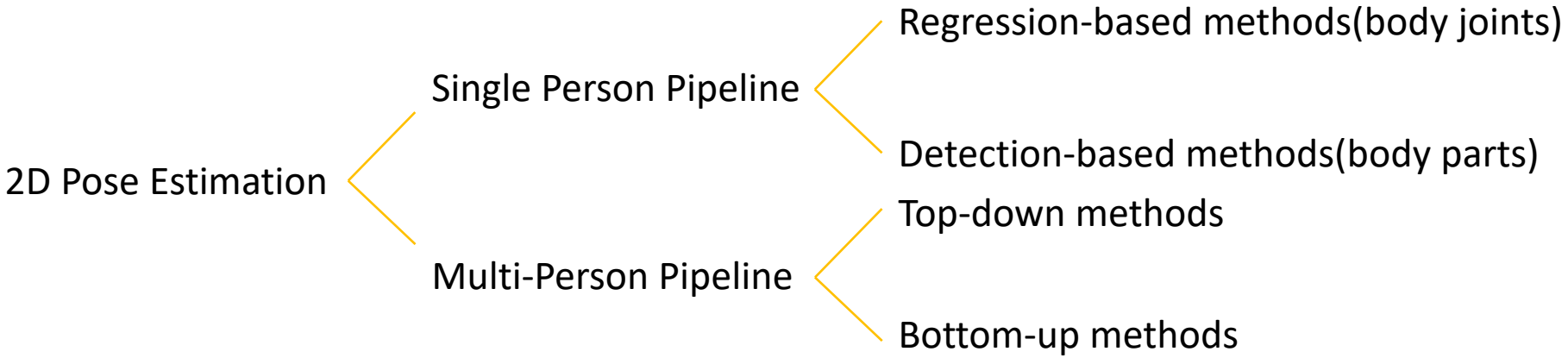
Background

- Human Pose Estimation



Background

- 2D Human Pose Estimation

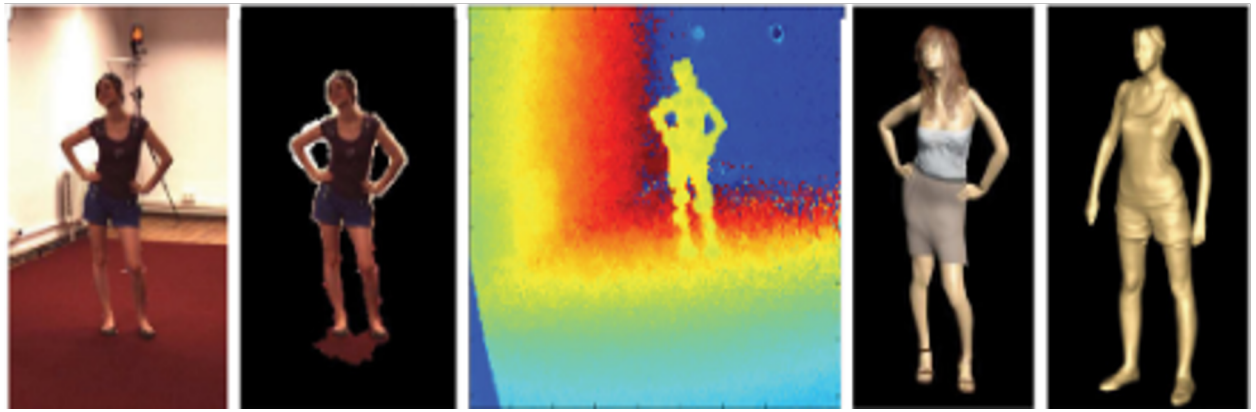


Single Person(DeepPose)

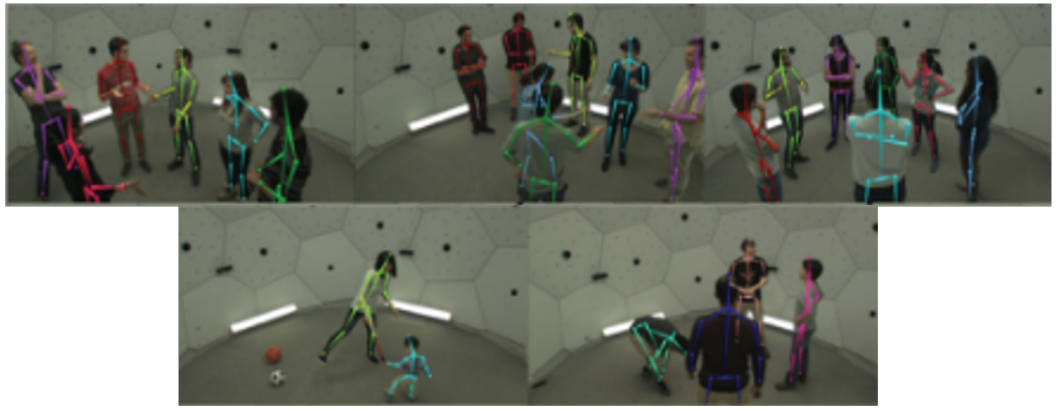
Multi Persons(OpenPose)

Background

- 3D Human Pose Estimation



Single Person(Human 3.6M)



Multi Persons(Panoptic)

- 3D Pose Estimation
  - Single Person Pipeline
    - Model-free methods
    - Model-based methods
  - Multi-Person Pipeline

Background

- 2D Human Pose Estimation

2D Pose Estimation
- Single Person Pipeline
  - Regression-based methods(body joints)
  - Detection-based methods(body parts)
- Multi-Person Pipeline
  - Top-down methods
  - Bottom-up methods

Matching Problem

Single Person(DeepPose)

Multi Persons(OpenPose)

Background

- 2D Human Pose Estimation

2D Pose Estimation
- Single Person Pipeline
  - Regression-based methods(body joints)
  - Detection-based methods(body parts)
- Multi-Person Pipeline
  - Top-down methods
  - Bottom-up methods

+ High Resolution

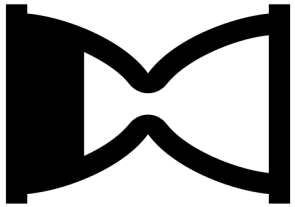
Matching Problem ➡ Solving Problem

Single Person(DeepPose)

Multi Persons(OpenPose)

Background

- Previous researches of Human Pose Estimation
  - Stacked hourglass network [CVPR 2016]
    : End-to-End method, Feeding the output of one as input into the next



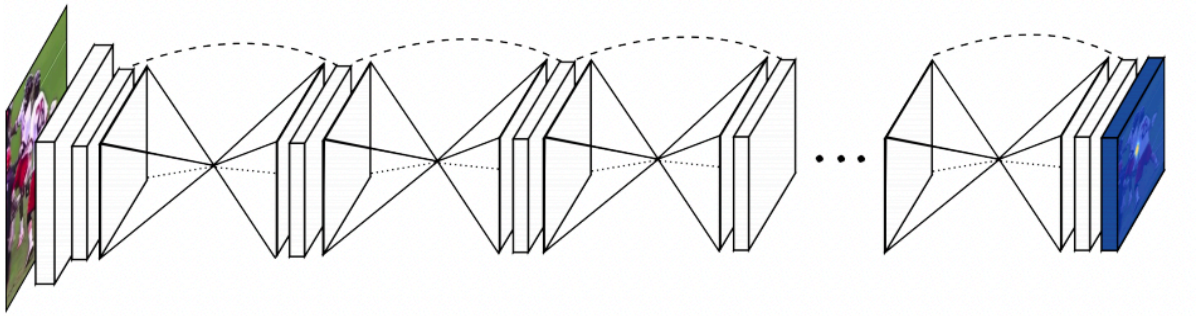


Fig. 1. Our network for pose estimation consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference.

Fig. 2. Example output produced by our network. On the left we see the final pose estimate provided by the max activations across each heatmap. On the right we show sample heatmaps. (From left to right: neck, left elbow, left wrist, right knee, right ankle)

Background

- Previous researches of Human Pose Estimation
  - Cascaded pyramid network [CVPR 2017]
    : Top-Down method, Estimate a pose inside the Bounding Box after detecting a person in image

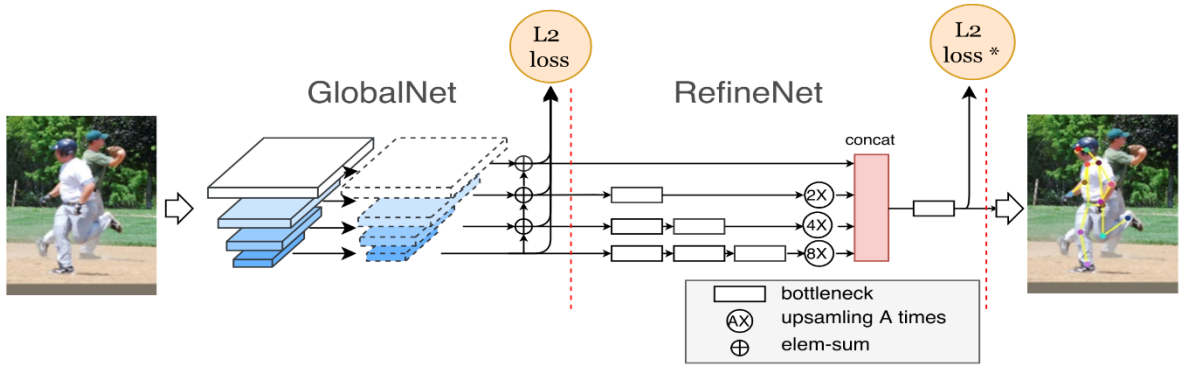


Figure 1. Cascaded Pyramid Network. "L2 loss*" means L2 loss with online hard keypoints mining.

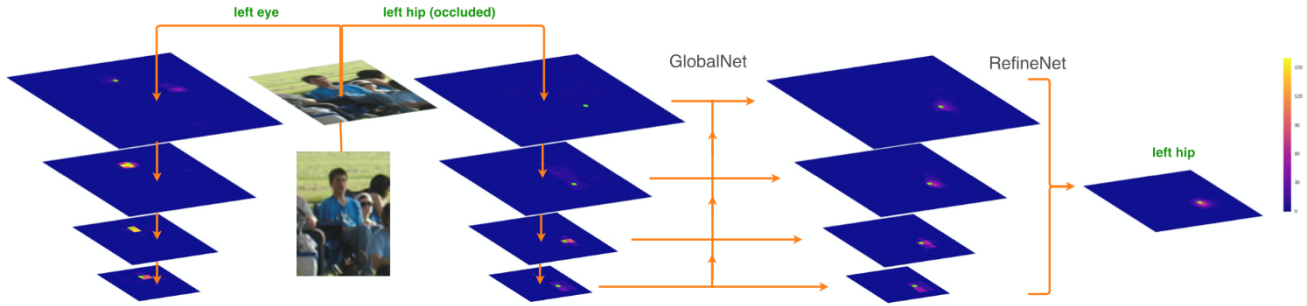
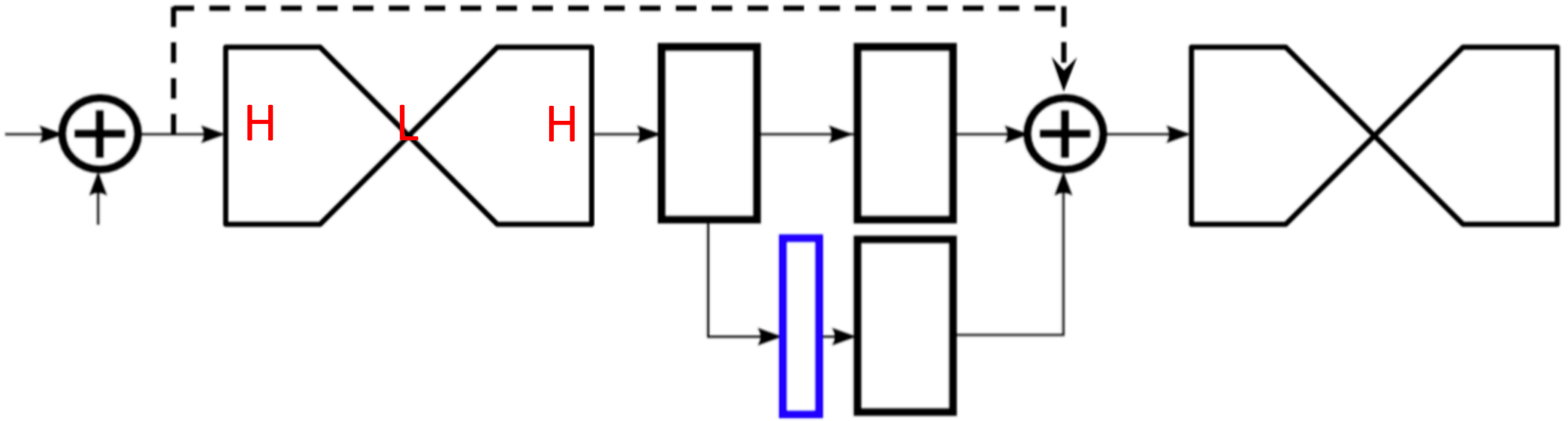


Figure 2. Output heatmaps from different features. The green dots means the groundtruth location of keypoints.

Related Works

- Stacked hourglass network [CVPR 2016]
  - High-to-Low and Low-to-High
  - Intermediate supervision



Intermediate supervision

Contribution

- **Deep High-Resolution Representation Learning for Human Pose Estimation**
  - Perform repeated multi-scale fusions to boost the high-resolution representations
    1) Connect high-to-low resolution subnetworks in parallel
    2) Repeated multi-scale fusions
    3) SOTA(COCO, MPII, PoseTrack)

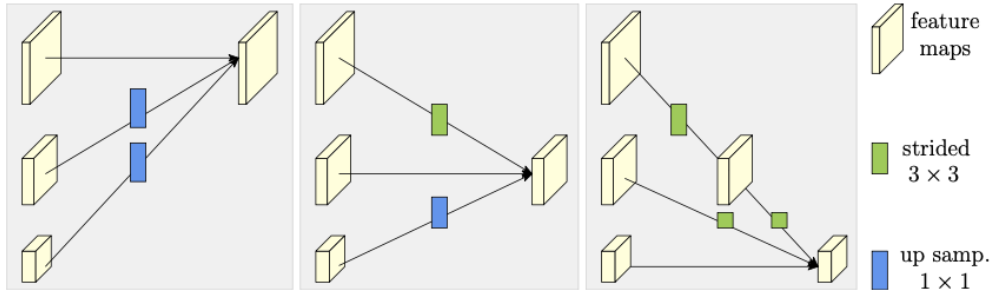


Figure 3. Illustrating how the exchange unit aggregates the information for high, medium and low resolutions from the left to the right, respectively. Right legend: strided 3×3 = strided 3×3 convolution, up samp. 1×1 = nearest neighbor up-sampling following a 1×1 convolution.

$$\begin{matrix} \mathcal{C}_{31}^{1} & \searrow & & \nearrow & \mathcal{C}_{31}^{2} & \searrow & & \nearrow & \mathcal{C}_{31}^{3} & \searrow & \\ \mathcal{C}_{32}^{1} & \rightarrow & \mathcal{E}_{3}^{1} & \rightarrow & \mathcal{C}_{32}^{2} & \rightarrow & \mathcal{E}_{3}^{2} & \rightarrow & \mathcal{C}_{32}^{3} & \rightarrow & \mathcal{E}_{3}^{3} \\ \mathcal{C}_{33}^{1} & \nearrow & & \searrow & \mathcal{C}_{33}^{2} & \nearrow & & \searrow & \mathcal{C}_{33}^{3} & \nearrow & \end{matrix}$$