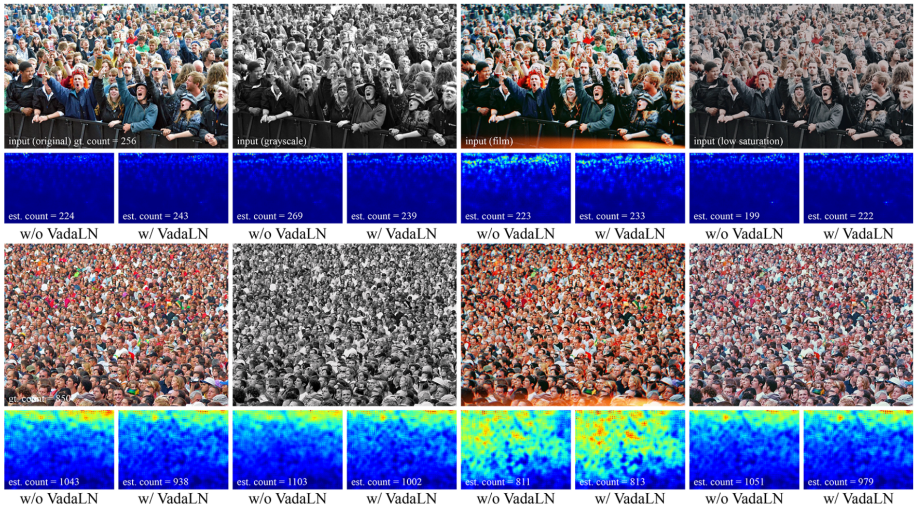
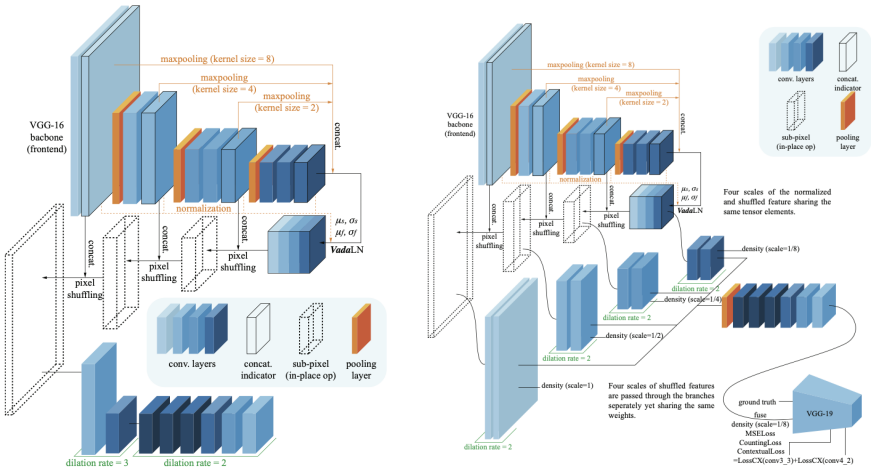
VGG-Embedded Adaptive Layer-Normalized Crowd Counting Net with Scale-Shuffling Modules

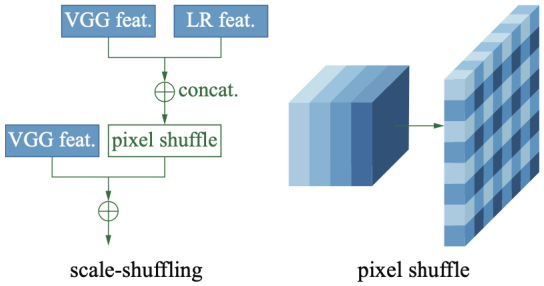
Dewen Guo, Jie Feng, Bingfeng Zhou
Peking University, CHN



Crowd counting is widely used in real-time conges- tion monitoring and public security. Due to the limited data, many methods have little ability to be generalized because the differences between feature domains are not taken into consideration. We propose VGG-embedded adaptive layer normalization (VadaLN) to filter the features that irrelevant to the counting tasks in order that the counting results should not be affected by the image quality, color or illumination. VadaLN is implemented on the pretrained VGG-16 backbone. There is no additional learning parameters required through our method. VadaLN incoporates the proposed scale-shuffling modules (SSM) to relax the distortions in upsampling operations. Besides, non- aligned training methdology for the estimation of density maps is leveraged by an adversarial contextual loss (ACL) to improve the counting performance.

We propose VGG-embedded adaptive layer normalization (VadaLN) to address this issue. Moreover, VadaLN also plays a role as a negative-feature filter to generalize the model to handle the images captured in the wild. Adaptive affine transformations using the specific layers in VGG-16 backbone is proposed to adapt LN to arbitrarily given cases of crowd counting. We formulate VadaLN to filter the image styles while keep the content feature for density estimation. Intuitively, different cases contain different illumination styles. Consid- ering of the style-representation layers in pretrained VGG- 16, in order to eliminate the influence from such layers while preserving the content of the crowd, the means and variances of these layers are precomputed for the normalization of the content layers.





Shuffle to the Desired Scale: With the operation mentioned above, to the specific task of crowd counting, we firstly resize the input images. Thus, the downscaled feature sizes in the ReLU layers of the VGG backbone can match those of the sub-pixel convoluted features.

Contextual loss is used to tackle the optimization problems for non-aligned data. The scheme is utilized for those source and target that are not necessarily aligned. With the ground truth generation by the Gaussian kernel, the pixel- wise difference between annotated maps can be considerably high. Thus, simply using pixel-wise loss may lead to the overfitting problem.

$$\mathrm{CX}(x,y) = \mathrm{CX}(X,Y) = \frac{1}{N}\sum_j \max_i \mathrm{CX}_{ij}$$

$$\mathcal{L}_{\mathrm{CX}}(x,y,l) = -\log\left(\mathrm{CX}\left(\Phi^l(x), \Phi^l(y)\right)\right)$$

$$\begin{aligned} ACL(x,y) = &-\lambda_1 \log\left(\mathrm{CX}\left(\Phi^{\mathrm{conv}_{3\_3}}(x), \Phi^{\mathrm{conv}_{3\_3}}(y)\right)\right) \\ &-\lambda_2 \log\left(\mathrm{CX}\left(\Phi^{\mathrm{conv}_{4\_2}}(x), \Phi^{\mathrm{conv}_{4\_2}}(y)\right)\right) \\ &+\lambda_3 \mathbb{E}_{x\sim p_{\mathrm{pred.}}}\left[\log D^*_{ACL}(x)\right] \\ &+\lambda_3 \mathbb{E}_{x\sim p_{gt}}\left[\log\left(1 - D^*_{ACL}(x)\right)\right] \end{aligned}$$