PAPER REVIEW: LARGE POSE 3D FACE RECONSTRUCTION FROM A SINGLE IMAGE VIA DIRECT VOLUMETRIC CNN REGRESSION

ESTUDIANTES:
LÓPEZ CÁCERES, JORGE ROBERTO
MAURICIO CONDORI, MANASSES ANTONI

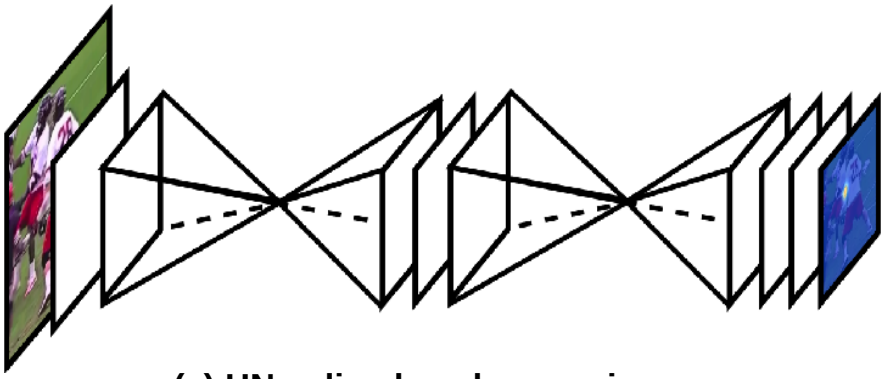
PROFESOR:
DSC MANUEL EDUARDO LOAIZA FERNÁNDEZ
IMÁGENES

UNIVERSIDAD CATÓLICA DE SAN PABLO
MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

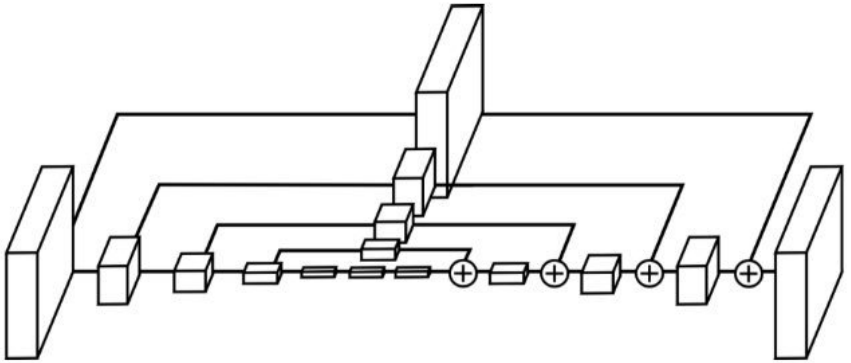
01 DE MARZO, 2018

Universidad Católica San Pablo
Centro de Investigación e Innovación en Ciencia de la Computación
CONCYTEC
CONSEJO NACIONAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN TECNOLÓGICA
CIENCIACTIVA
Becas y Co-financiamiento de Concytec

HOURGLASS NETWORKS



(a) HN aplicada sobre una imagen



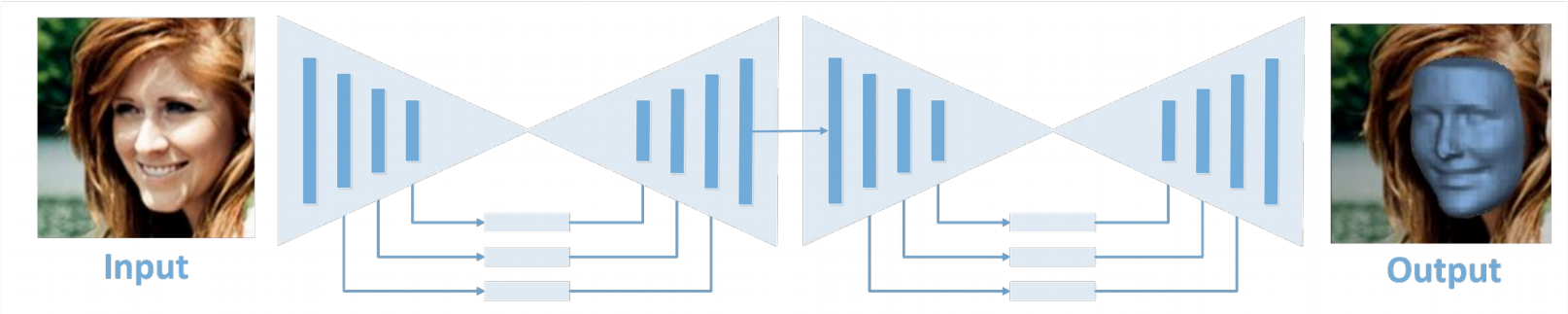
(b) Modelo de red residual

(c) A la izquierda vemos la postura final estimada proporcionada por las activaciones máximas en cada mapa de calor. A la derecha las salidas de la red

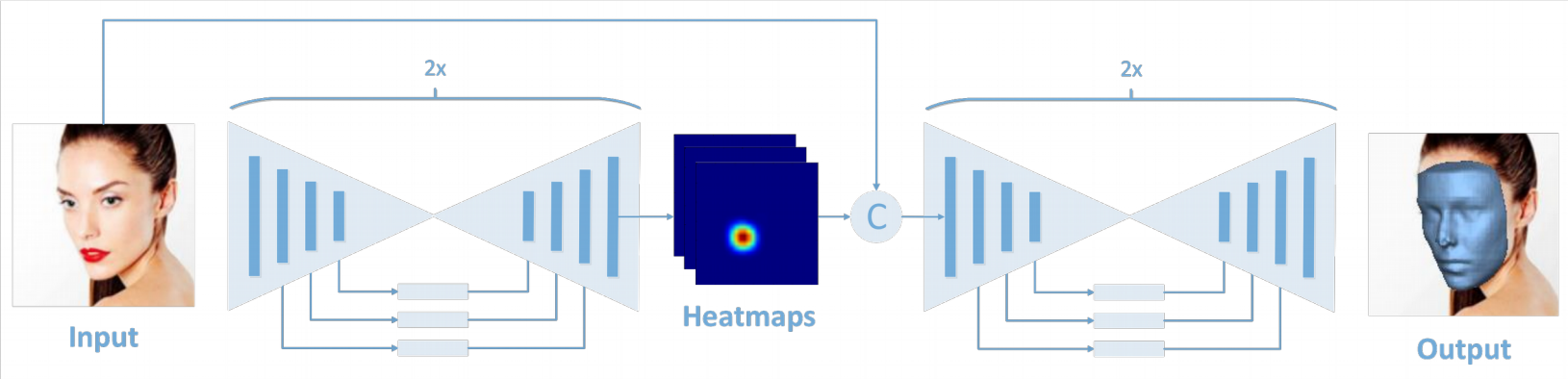
Stacked Hourglass Networks for Human Pose Estimation [2]

[2] Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." European Conference on Computer Vision. Springer, Cham, 2016.

ARQUITECTURAS PROPUESTAS I

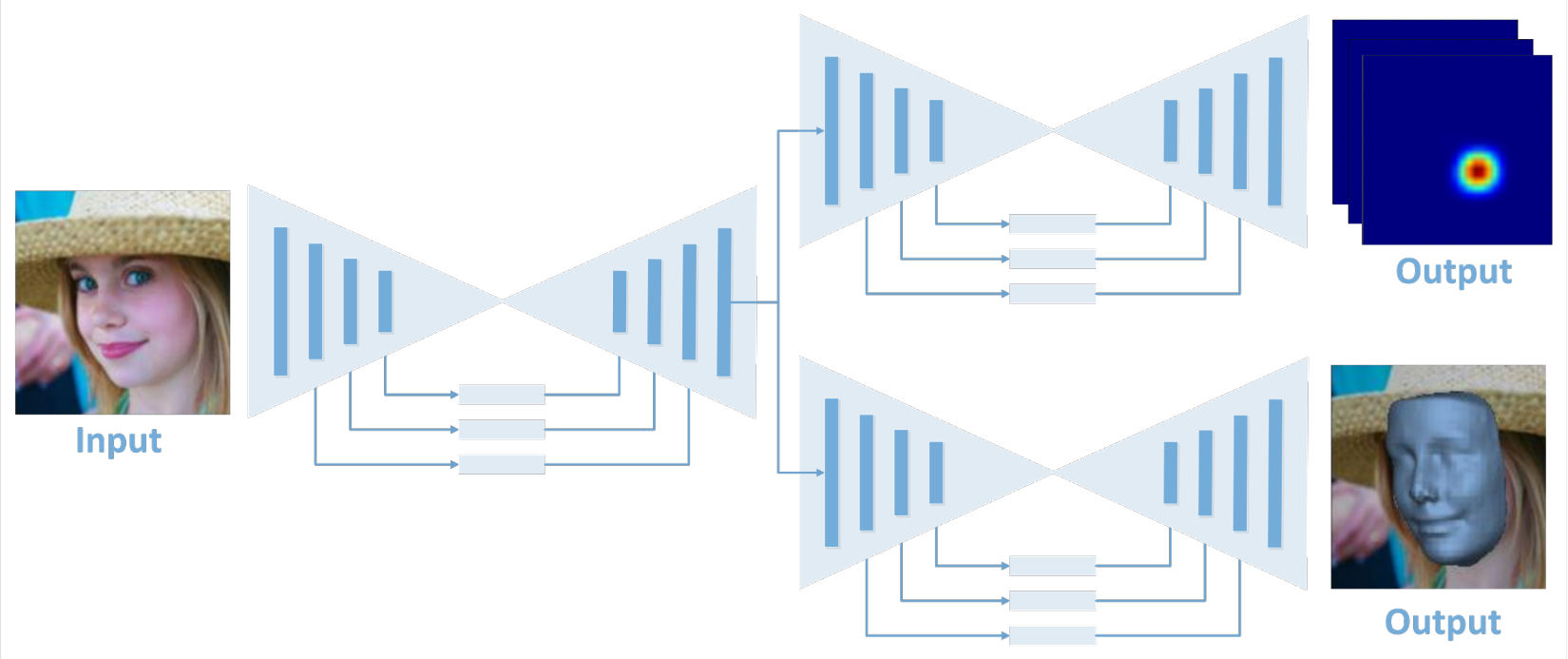


(a) VRN acepta como entrada una entrada RGB y regresa directamente un volumen 3D evitando por completo el ajuste de un 3DMM. Cada rectángulo es un módulo residual de 256 características.



(a) VRN-guided primero detecta la proyección 2D de los puntos de referencia en 3D y los apila con la imagen original. Esta pila se alimenta a la red de reconstrucción, que directamente regresa el volumen.

ARQUITECTURAS PROPUESTAS II



(c) VRN-multitask regresa tanto el volumen facial 3D como un conjunto de puntos faciales escasos.

ENTRENAMIENTO

METHOD	DESCRIPTION
Learning Algorithm	RMSProp
learning rate	10e-4 / 10e-5
Data Augmentation	Random augmentation: - Rotation/Translation/ Scaling. - 20% flipped
Cost Function	Sigmoid cross entropy loss function
Input	2D images
Output	- 3D Volume - Scan Images

(Tabla 1) Parámetros del Entrenamiento

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2$$
$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t$$

RMSProp

$$l_1 = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D} W_{whd} log(W'_{whd}) + (1 - W_{whd}) log(1 - W'_{whd}) \quad (1)$$

Where W'_{whd} is the corresponding sigmoid output at voxel w, d, h of the regressed volume.

$$NME = \frac{1}{N}\sum_{k=1}^{N}\frac{||x_k - y_k||_2}{d} \quad (2)$$

where N is the number of vertices per facial mesh, d is the 3D interocular distance and x_k ,y_k are vertices of the grouthtruth and predicted meshes. The error is calculated on the face region only on approximately 19,000 vertices per facial mesh.

Each of our architectures was trained end-to-end using RMSProp with an initial learning rate of 10^{-4} , which was lowered after 40 epochs to 10^{-5} . During training, random augmentation was applied to each input sample (face image) and its corresponding target (3D volume): we applied in-plane rotation $r \in [-45°, ..., 45°]$, translation $t_z, t_y \in [15, ..., 15]$ and scale $s \in [0.85, ..., 1.15]$ jitter. In 20% of cases, the input and target were flipped horizontally. Finally, the input samples were adjusted with some colour scaling on each RGB channel.

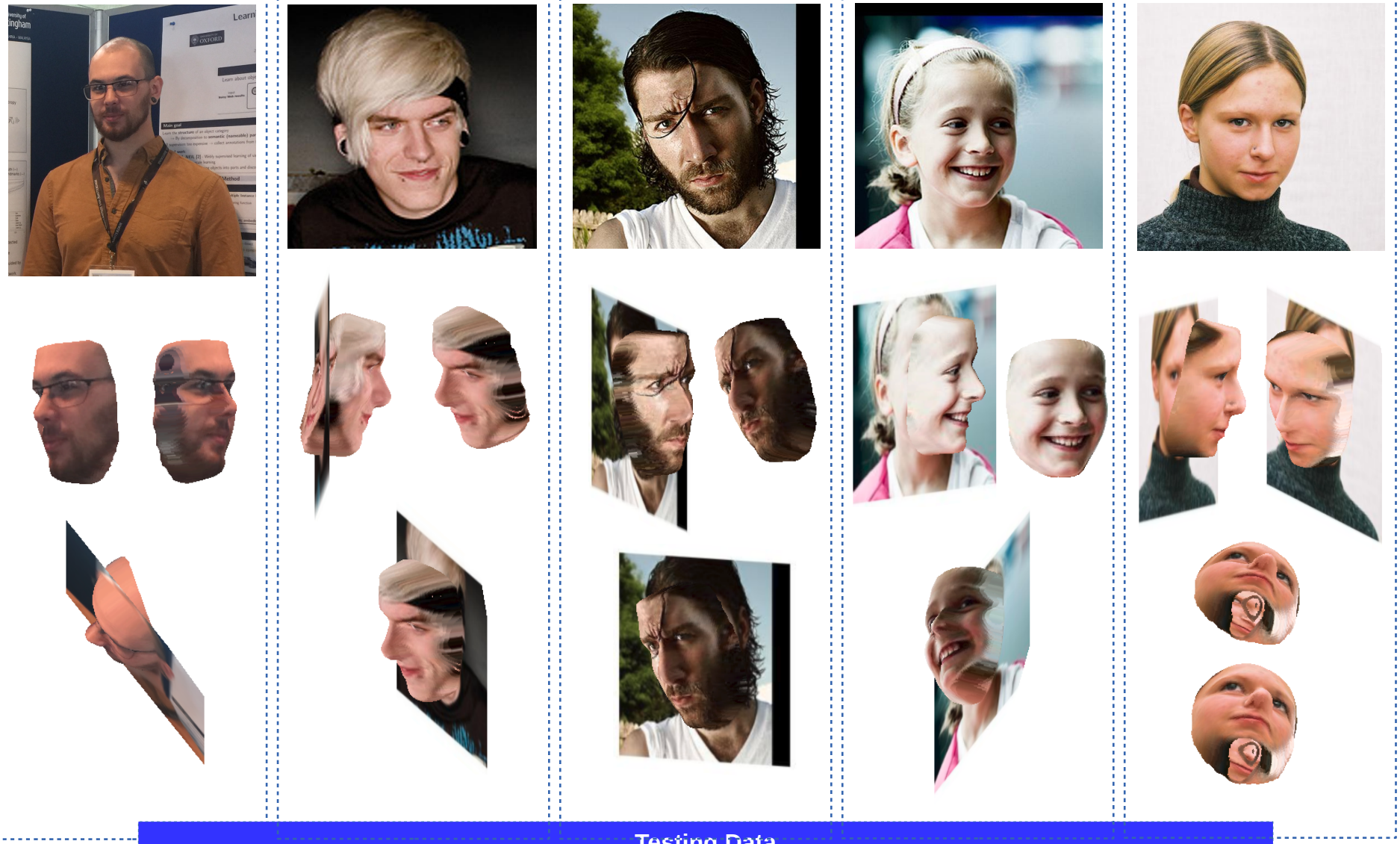
In the case of the VRN - Guided, the landmark detection module was trained to regress Gaussians with standard deviation of approximately 3 pixels ($\sigma = 1$).

COMPARATIVO – ESTADO DEL ARTE

Method	AFLW2000	BU4DFE	Florence
VRN	0.0676	0.0600	0.0568
VRN - Multitask	0.0698	0.0625	0.0542
VRN - Guided	**0.0637**	**0.0555**	**0.0509**
3DDFA [28]	0.1012	0.1227	0.0975
EOS [7]	0.0971	0.1560	0.1253

(Tabla 2) Resultados Comparativos

RESULTADOS



Testing Data

CASOS ERRONEOS

Our Data