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

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

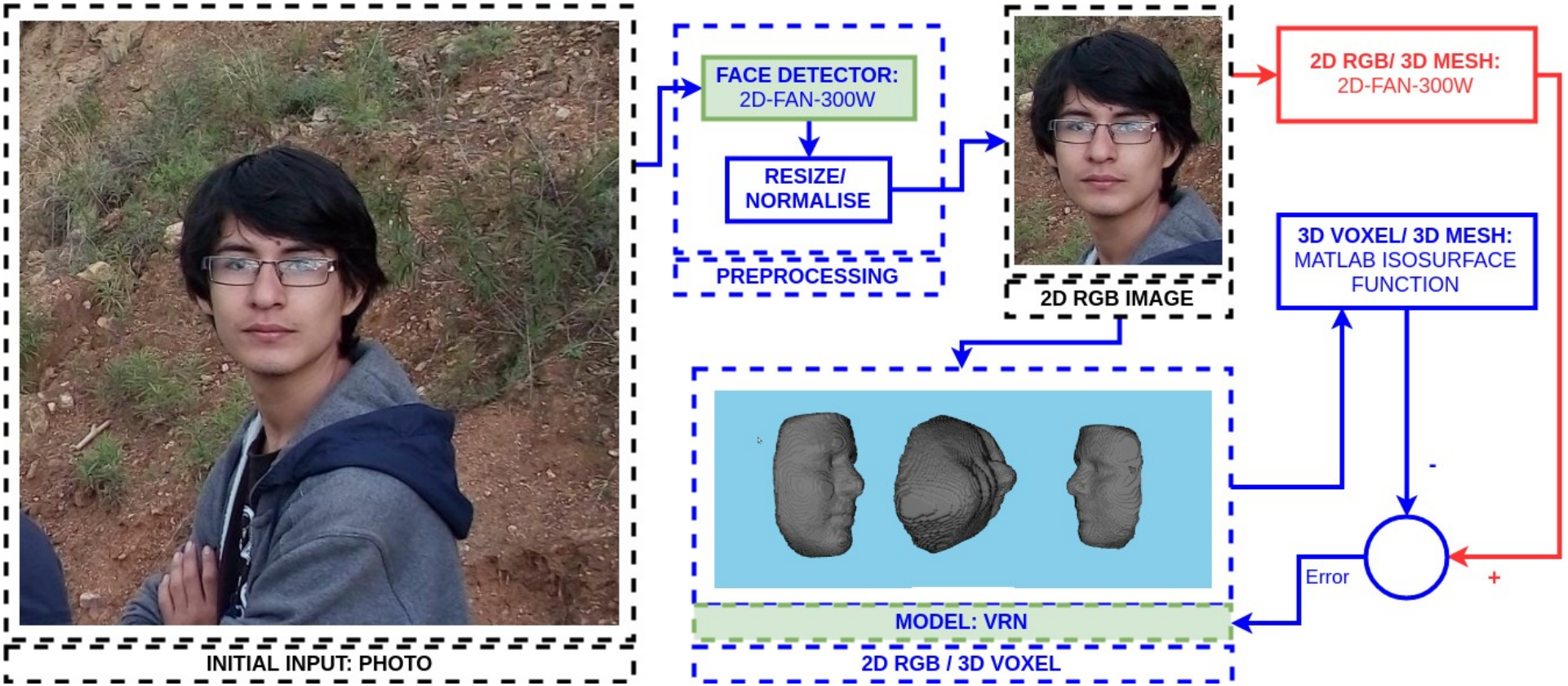
PROFESSOR:
DSC MANUEL EDUARDO LOAIZA FERNÁNDEZ

UNIVERSIDAD CATÓLICA DE SAN PABLO
MASTER OF COMPUTER SCIENCE

MARCH 22, 2018

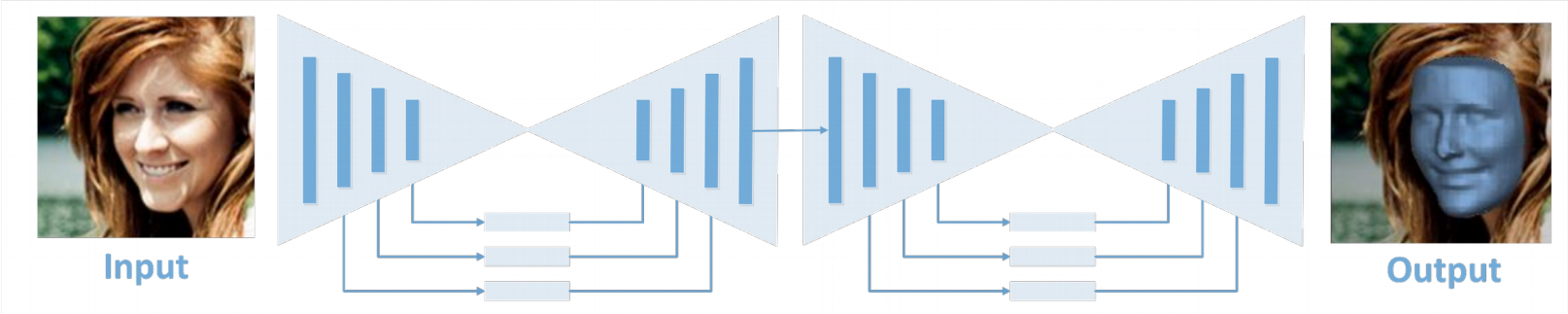
Universidad Católica San Pablo
Centro de Investigación e Innovación en Ciencia de la Computación
CONCYTEC CONSEJO NACIONAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN TECNOLÓGICA
CIENCIACTIVA Becas y Co-financiamiento de Concytec

GENERAL PIPELINE

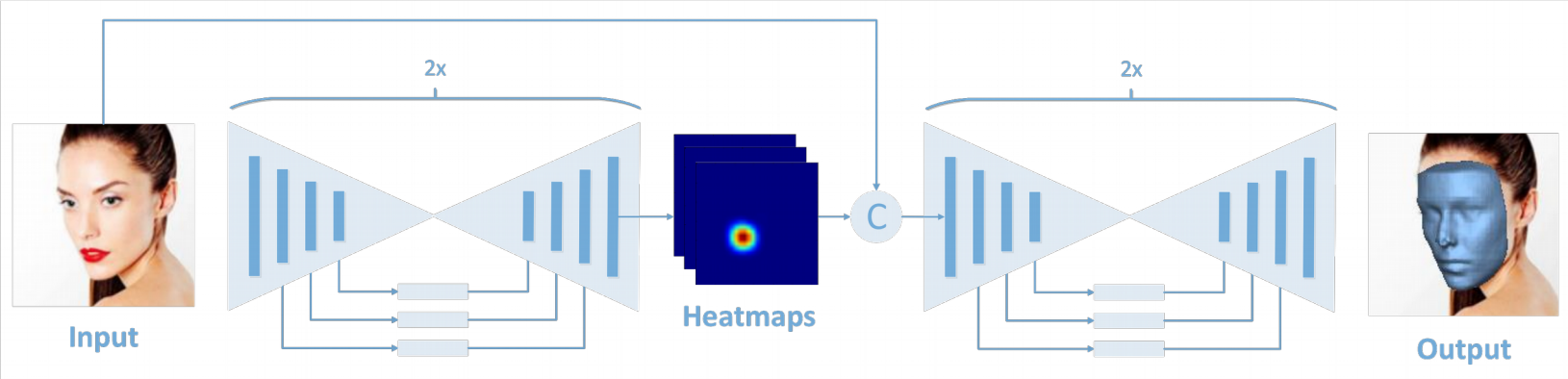


[1] Jackson, Aaron S., et al. "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.
[2] Peng, Xi, et al. "Reconstruction-based disentanglement for pose-invariant face recognition." intervals 20 (2017): 12.

PROPOSED ARCHITECTURES I



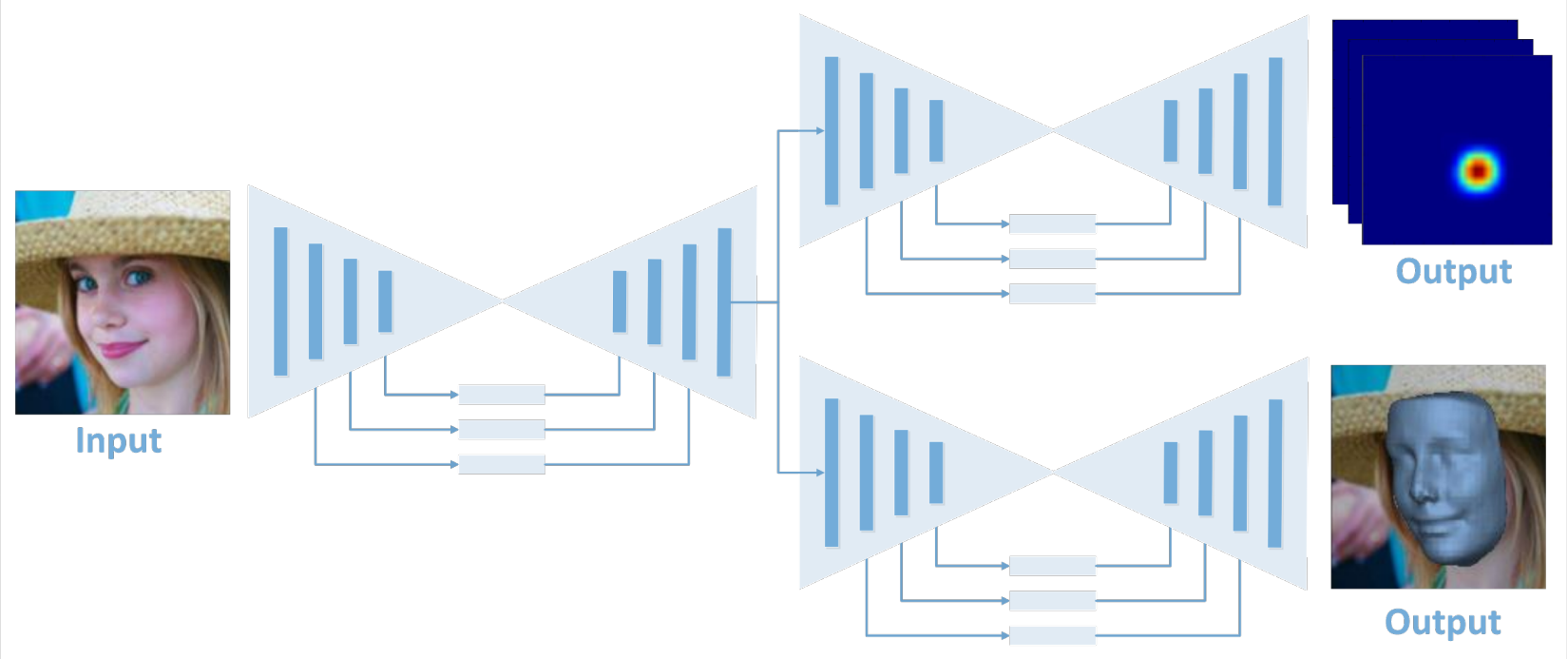
(a) Volumetric Regression Network (VRN) accepts as input an 2D RGB input and directly regresses a 3D volume completely bypassing the fitting of a 3DMM. Each rectangle is a residual module of 256 features.



(b) VRN - Guided architecture firsts detects the 2D projection of the 3D landmarks, and stacks these with the original image. This stack is fed into the reconstruction network, which directly regresses the volume.

[2] Jackson, Aaron S., et al. "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.

PROPOSED ARCHITECTURES II



(c) VRN - Multitask architecture regresses both the 3D facial volume and a set of sparse facial landmarks.

[2] Jackson, Aaron S., et al. "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.

TRAINING DESCRIPTION

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 RGB image
Output	- 3D 192 × 192 × 200 voxel 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 ($\sigma = 1$).

[2] Jackson, Aaron S., et al. "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression." Computer Vision (ICCV), 2017 IEEE International Conference on. IEEE, 2017.

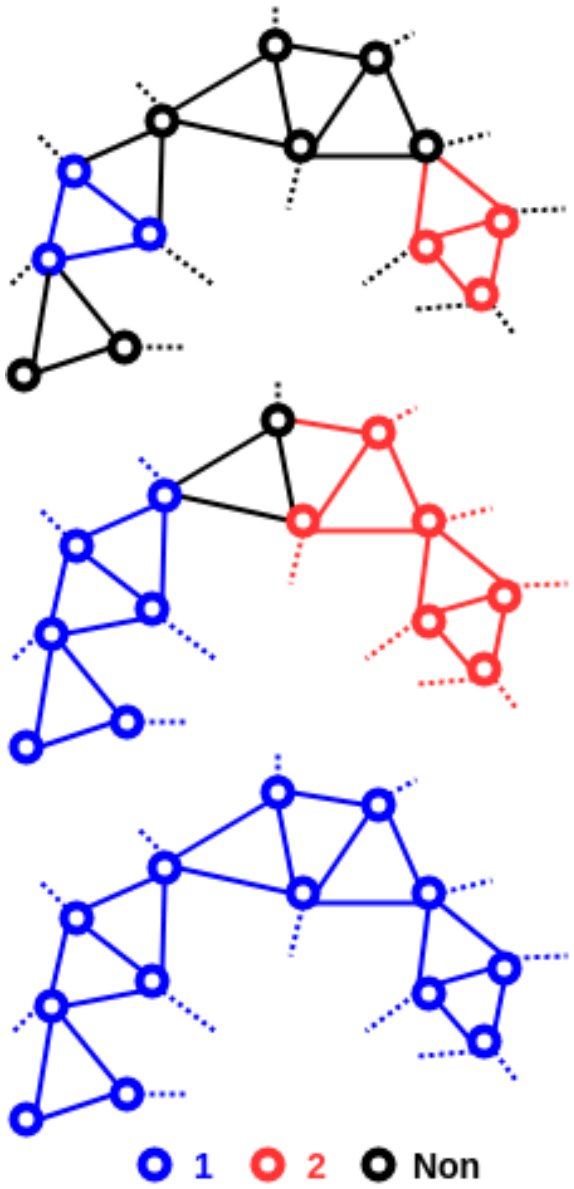
PREVIOUS RESULTS

Testing Data

PREVIOUS RESULTS II

Testing Data

DYNAMIC LABELING BASED ON NEIGHBORHOOD

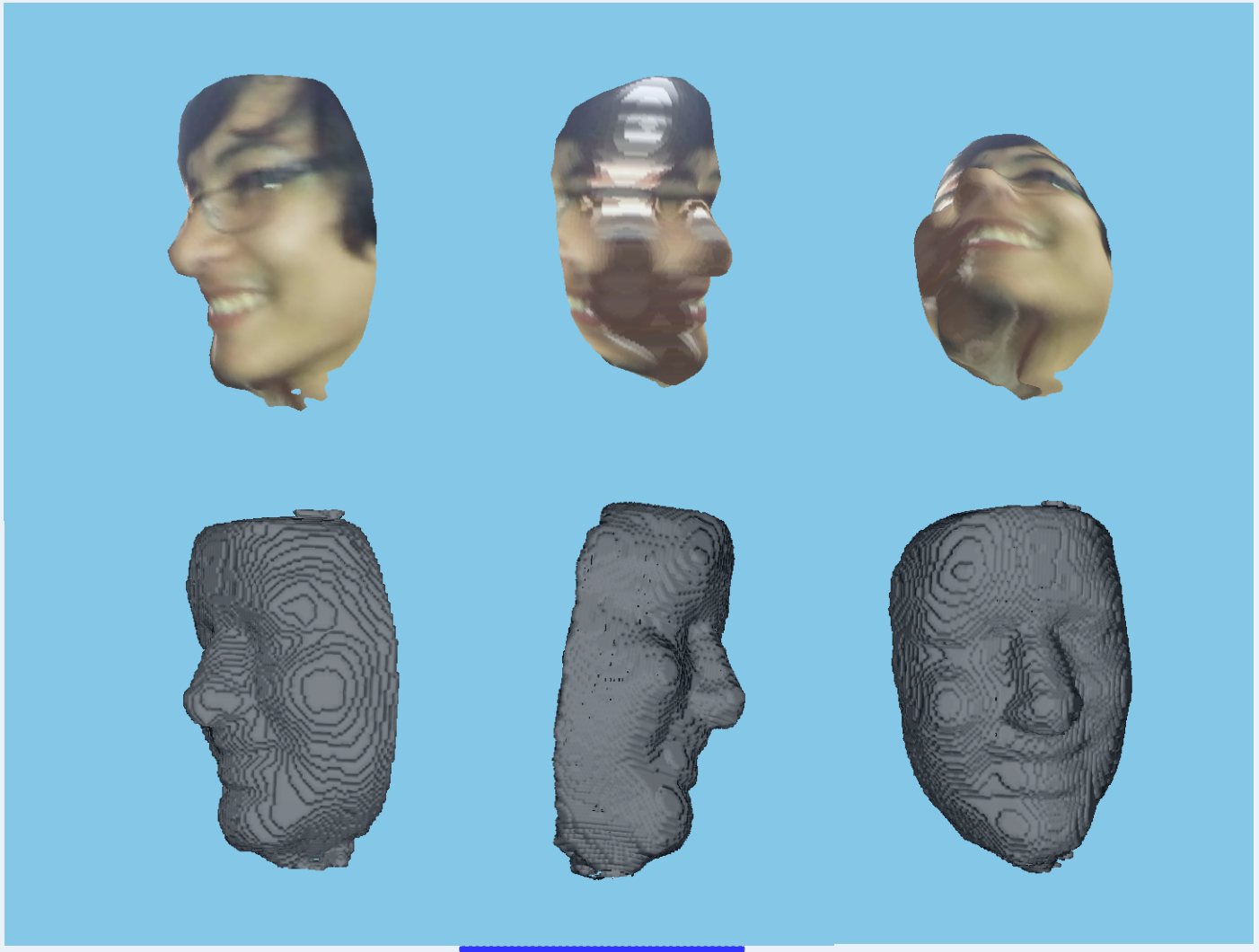


Algorithm Object Detection Algorithm (Mesh):
Mesh fragments → graph point.
for each fragment **do**
  if at least one vertex has a label: **then**
    Search all labels related to the vertices labels.
    From all labels look for the smallest number → m.
    Previous labels are now connected with m.
    label the vertices with m
  else
    Create new label g.
    Label related vertices with g.
  end if
  for each label **do**
    Find the first parent generated in tree.
    Relate directly with the first father.
    if Child does not exist **then**
      Create it.
    end if
  end for
  Label each vertex with their own first father.
  Calculate objects size.
  Preserve just the biggest object.
end for
return Mesh out

PSEUDOCODE

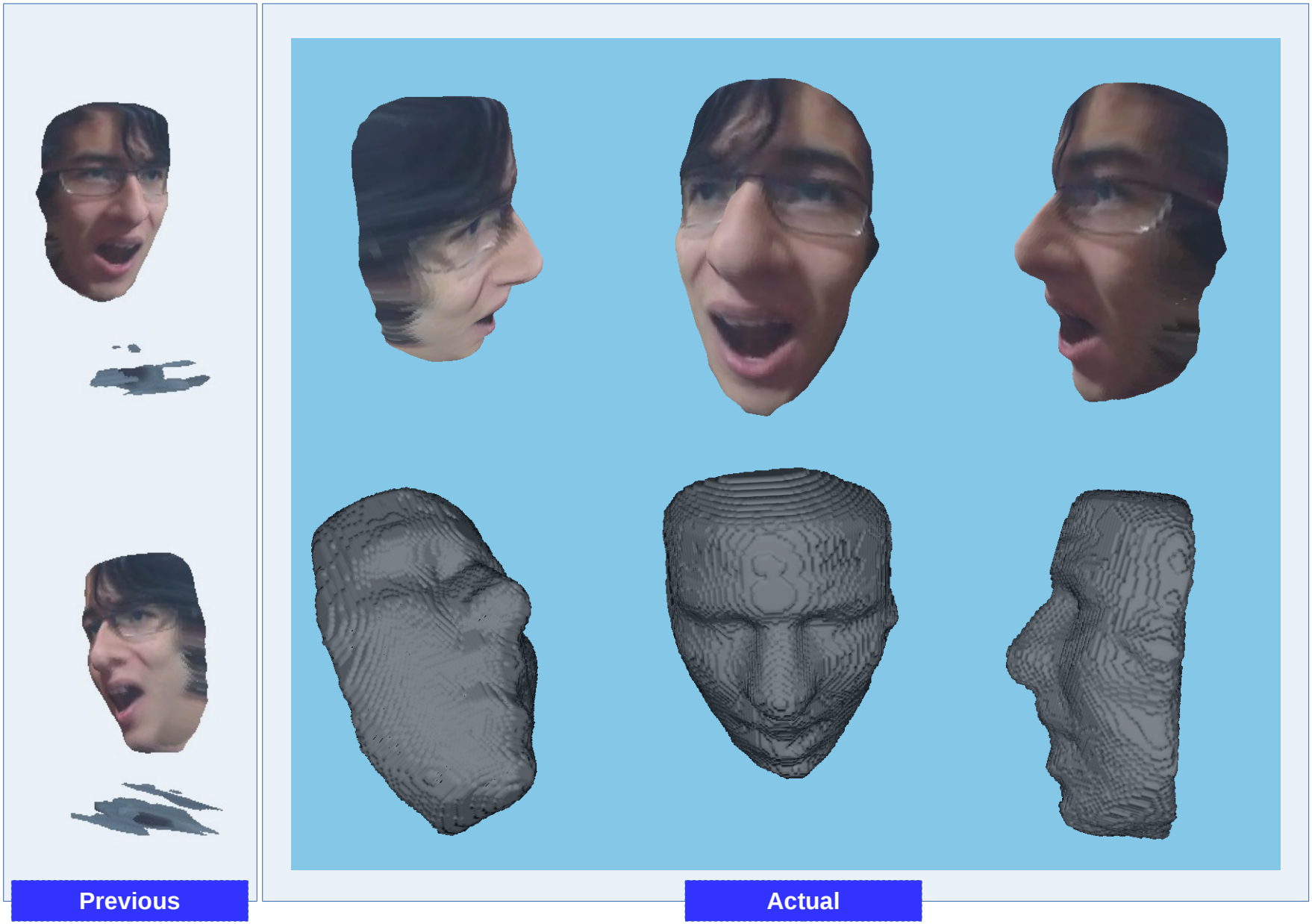
WHAT WE ACHIEVED I

Previous



Actual

WHAT WE ACHIEVED II



CONCLUSIONS

- Posprocessing is better in our case due target generation function and feedback error functions work together, which means the error is implicit.
- Residual Network is very usefull to avoid vanishing gradient problem, but also implies wheter error will maintein over epoch till convergence.
- VRN-Guided should be better than VRN presented but the article does not give us enought implementation details to reproduce it, also heatmaps to guide the network to find out points of interest reduce the problem complexity.