Supplementary Material of IEEE FG'19 Paper: Extended Gaze Following: Detecting Objects in Videos Beyond the Camera Field of View

Benoit Massé[1], Stéphane Lathuilière[1,2], Pablo Mesejo[1,3] and Radu Horaud[1]
[1] Inria & Univ. Grenoble Alpes, France, [2] University of Trento, Italy, [3] University of Granada, Spain

ABLATION STUDY: T

We report experiments to measure the impact in performance of the sequence length T in Fig. 1. Precisely, we selected *Mean-2D-Enc* (as best model on *Vernissage*) and *3D/2D U-Net* (as best model on *synthetic*) and compute the *f1-score* evolution for these two networks varying T from 10 to 450. Both networks behave similarly to the results reported before: *3D/2D U-Net* is consistently better on *synthetic* data than *Mean-2D-Enc*, and consistently worse on the *Vernissage* dataset. We observe that the performances of both networks tend to increase with the sequence length on *synthetic* data, though quite slowly for $T > 150$. However, when the networks are transfered to be used on the *Vernissage* dataset, the *f1-score* stops increasing past $T = 200$ or 250. Moreover, the variances are sometimes quite higher, which could indicate a more unstable training process. This validates the choice of $T = 200$ for our experiments.

OTHER SYNTHETIC EXAMPLES

Example of generated scenarios in Fig. 2-3-4. Fig. 2 is the generated scenario used in the paper.

This work is supported by ERC Advanced Grant VHIA #340113.

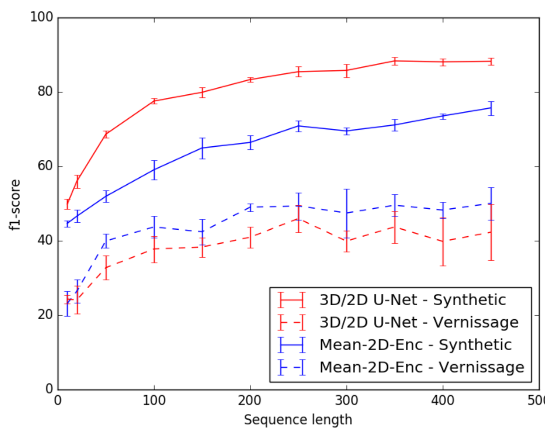
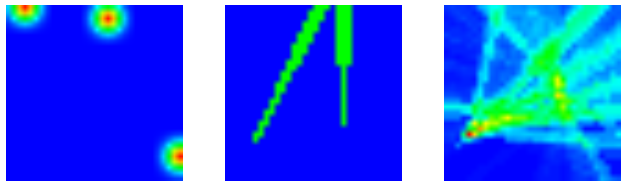
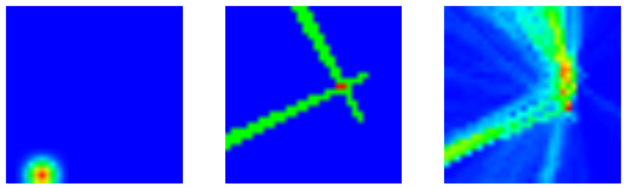


Fig. 1: Performance obtained on the *synthetic* and *Vernissage* datasets with RGB data. We measure the *f1-score* with different values of sequence length T.



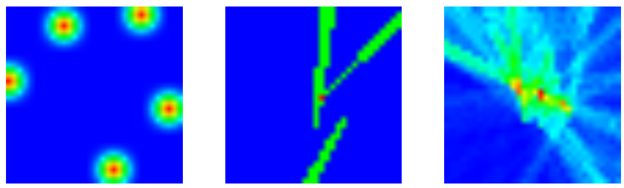
(a) *Object heat-map* (b) *Gaze heat-map* (c) Mean *gaze heat-map*

Fig. 2: Heat-maps from a synthetic scenario generated randomly, with 2 people ($N = 2$) and 3 objects ($M = 3$). (a): the ground truth *Object heat-map* $\mathbf{\Omega}$ used for training or evaluation. (b): a *Gaze heat-map* randomly chosen among the sequence. (c): the mean *gaze heat-map* over the sequence.



(a) *Object head-map* (b) *Gaze heat-map* (c) Mean *gaze heat-map*

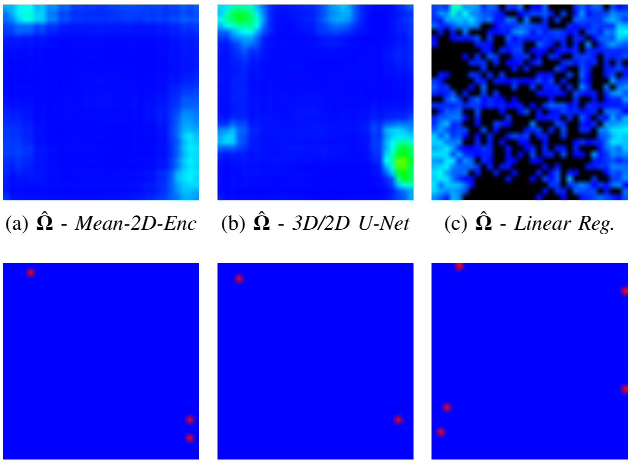
Fig. 3: Heat-maps from a synthetic scenario generated randomly, similar to Fig. 2, but with a different setup: 2 people ($N = 2$) and 1 object ($M = 1$).



(a) *Object head-map* (b) *Gaze heat-map* (c) Mean *gaze heat-map*

Fig. 4: Heat-maps from a synthetic scenario generated randomly, similar to Fig. 2, but with a different setup: 3 people ($N = 3$) and 5 objects ($M = 5$).

In Fig. 5, the predicted *gaze heat-maps* $\hat{\mathbf{\Omega}}$ for several learning-based approaches applied on the *synthetic* scenario from Fig. 2 are displayed. The architectures *Mean-2D-Enc* and *Linear Reg.* use the average *gaze heat-map* $\frac{1}{T}\sum_{t=1}^{T}\mathbf{\Gamma}_t$ as input, whereas *3D/2D U-Net* takes the whole concatenated sequence $\mathbf{\Gamma}_{1:T}$. Contrary to the experiments on the *Vernissage* dataset, We observe that the *3D/2D U-Net* yields an *object*



(a) $\hat{\mathbf{\Omega}}$ - *Mean-2D-Enc* (b) $\hat{\mathbf{\Omega}}$ - *3D/2D U-Net* (c) $\hat{\mathbf{\Omega}}$ - *Linear Reg.*

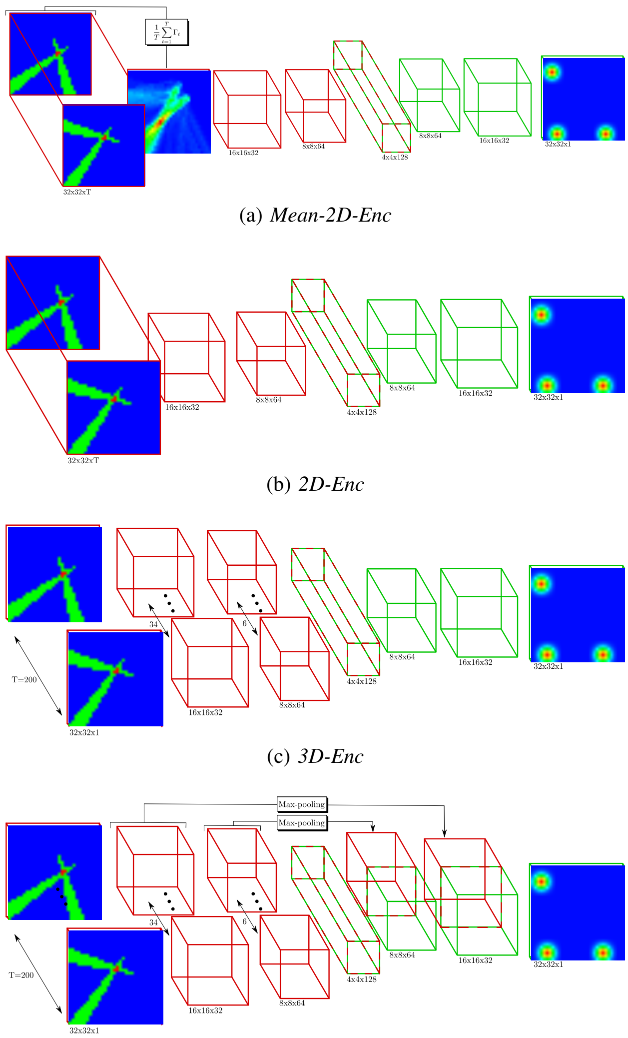
(d) Obj - *Mean-2D-Enc* (e) Obj - *3D/2D U-Net* (f) Obj - *Linear Reg.*

Fig. 5: Results of three methods on the *synthetic* sequence from Fig. 2 (a), (b), (c): Estimates $\hat{\mathbf{\Omega}}$ of the *synthetic object heat-map* $\mathbf{\Omega}$ from Fig. 2a using three different architectures. (d), (e), (f) : Corresponding objects positions, obtained as the highest local maxima from $\hat{\mathbf{\Omega}}$. Black pixels in (c) indicate negative values.

heat-map $\hat{\mathbf{\Omega}}$ closer to the expected one $\mathbf{\Omega}$ than the other models, and lead to a higher precision. This is consistent with the quantitative results reported in Table I in the main paper.

ARCHITECTURES

Fig. 6 is an illustration of the convolutional encoder/decoder architectures proposed in section III-B of the main paper.



(a) *Mean-2D-Enc*

(b) *2D-Enc*

(c) *3D-Enc*

(d) *3D/2D U-Net*

Fig. 6: Proposed convolutional encoder/decoder architectures