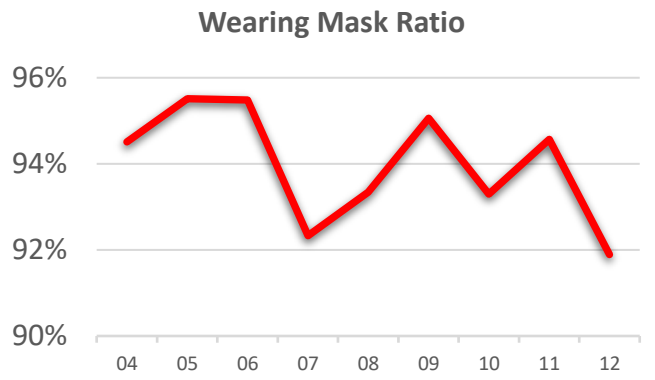
EFFECTIVENESS OF DETECTION-BASED AND REGRESSION-BASED APPROACHES FOR ESTIMATING MASK-WEARING RATIO

Khanh-Duy Nguyen[1], Trung-Nghia Le[2], Huy H. Nguyen[2], Junichi Yamagishi[2], Isao Echizen[2]
(1) University of Information Technology – VNUHCM
(2) National institute of Informatics

Overview

- **Motivation:**
  - Investigate **human behavior in COVID-19**: mask wearing, social distancing.
  - Collect **reliable statistics of people** that pass through **street cams.**

Percentage of people wearing a mask?



Captured from **Shibuya Crossing Live Camera** in March 2021 (08:00-14h00 every day)

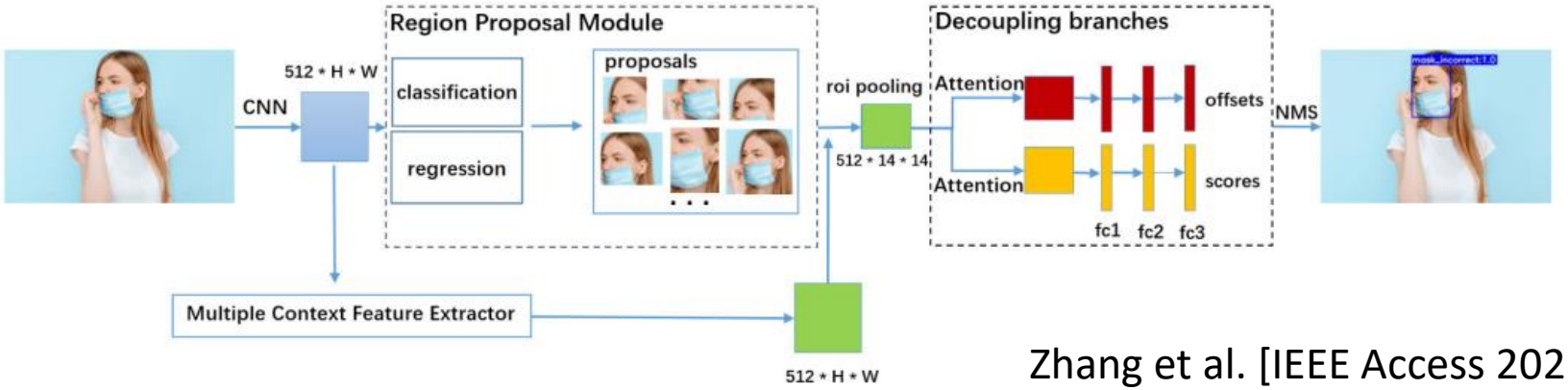
Overview

- **Our contributions:**

1. We introduce a **large-scale dataset** of images extracted from street-view videos for use in estimating the face mask ratio.
   - Our dataset contains 18,088 video frames with more than 580,000 face annotations.
   - To our best knowledge, this is the first face mask dataset containing images extracted from street-view videos.
2. We present a comparative evaluation of two approaches to estimating the mask-wearing ratio: the **detection-based** approach and the **regression-based** approach.
   - The advantages and disadvantages of the two approaches are also discussed.

Introduction:
Estimating the Mask-wearing Ratio

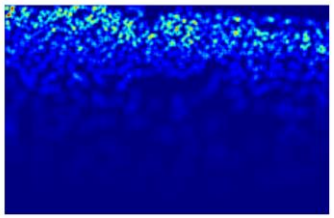
- **Detection-based approach:**
  - Pioneering works focused on detecting masked and unmasked faces in images.
    - They used an object detector, such as R-CNN or YOLO.
    - The obtained results can be further processed to make statistics and warnings.

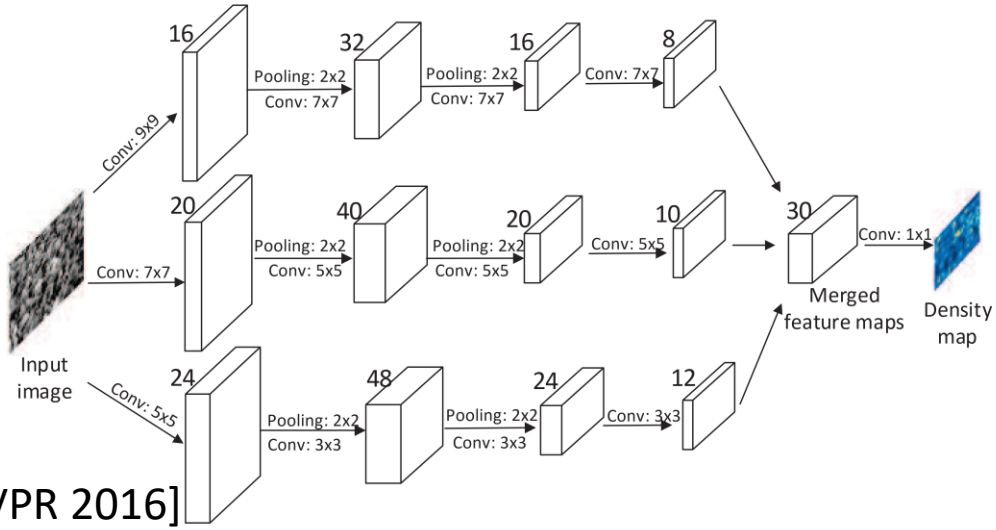


Zhang et al. [IEEE Access 2021]

Introduction: Estimating the Mask-wearing Ratio

- **Regression-based approach:**
  - It is also feasible to use other **crowd-counting approaches** to estimate the number of people wearing masks.
    - Recent CNN-based crowd-counting methods are especially efficient for congested pedestrian flows thanks to the utilization of density maps.
    - More recent efforts have taken a regression-based approach, using a Convolutional Neural Network to accurately and quickly predict density maps.





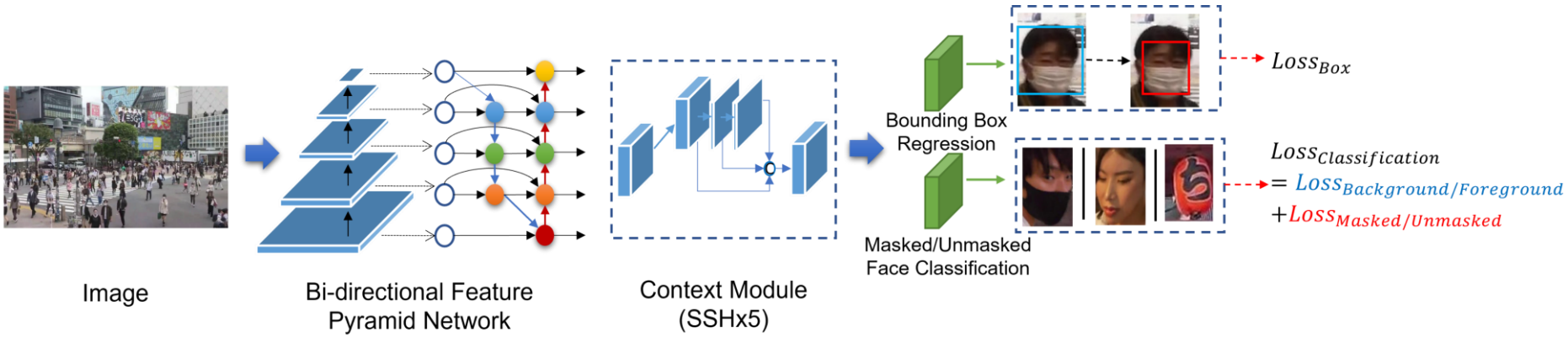
Zhang et al. [CVPR 2016]

Introduction:
Estimating the Mask-wearing Ratio

- **Detection-based approach vs. Regression-based approach:**
  - For crowd counting: Liu et al. [CVPR 2018]
    - The detection-based method counts people accurately in sparse scenes but has unreliability in congested scenes.
    - Meanwhile, the regression-based method usually overestimates the crowd in low-density areas.
  - For estimating the mask-wearing ratio
    - The effectiveness of regression-based methods on the mask-wearing ratio estimation has not been investigated.

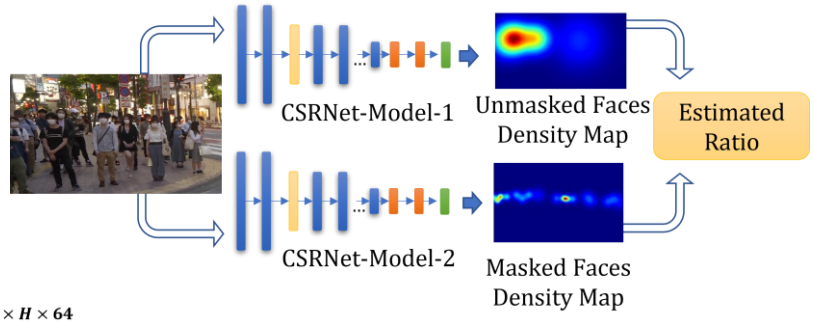
Our Proposed Methods

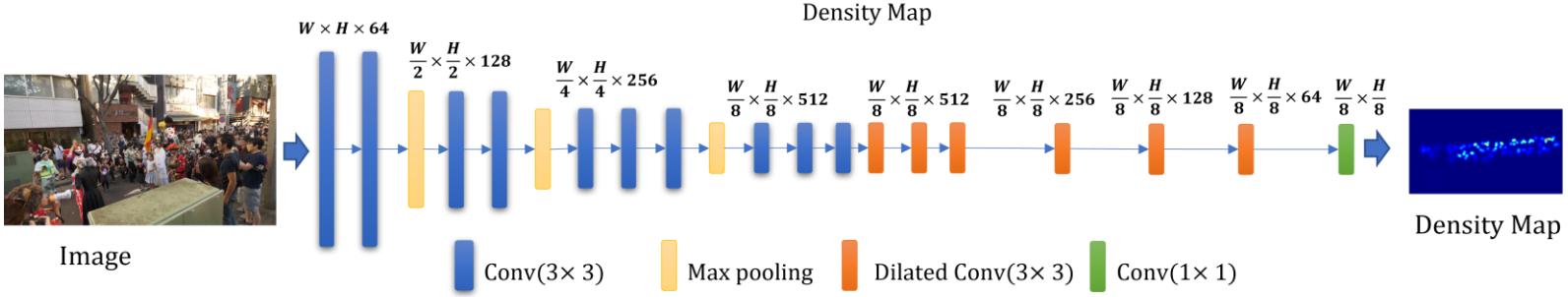
- **RetinaFace-based Mask-wearing Ratio Estimation:**
  - RetinaFace detector with enhanced components:
    - Mask/Unmasked Face Classification
    - Bi-FPN, Multi-task Loss
  - Ratio estimation: the mask-wearing ratio is simply calculated by dividing the number of masked faces by the total number of detected faces.



Our Proposed Methods

- **CSRNet-based Mask-wearing Ratio Estimation:**
  - Use two CSRNet models to separately predict the numbers: masked faces and unmasked faces.
  - Ratio estimation: we simply divide the number of unmasked faces by the total number of faces (masked and unmasked).





Large-scale NFM Dataset

- **Existing masked face datasets flaws[(1)(2)]:**
  - Many images were crawled from the Internet (webcams, newspapers, day-life photos).
  - Faces generally had **high resolution** and were in **frontal view**.
- **Our street camera video dataset:**
  - Images were crawled from street cameras.
  - Faces were **drastically smaller**, much **less clear**, and at **various angles.**

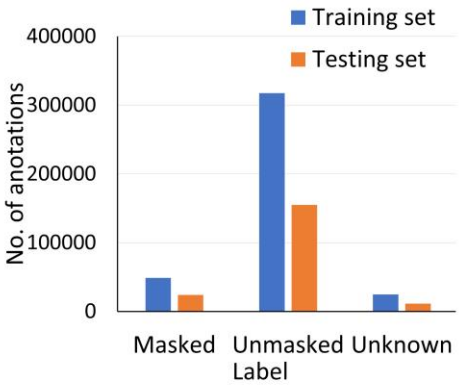

MAFA Dataset


Street Camera Video

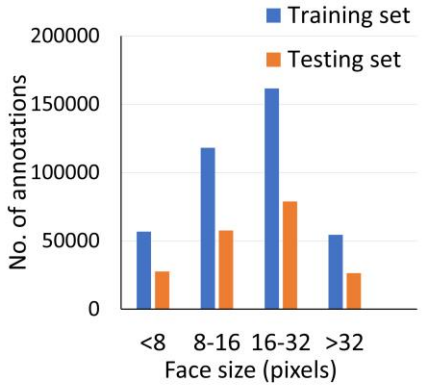
(1) Kaggle Medical Mask Dataset: Link
(2) MAFA - MAsked Faces: Link

Large-scale NFM Dataset

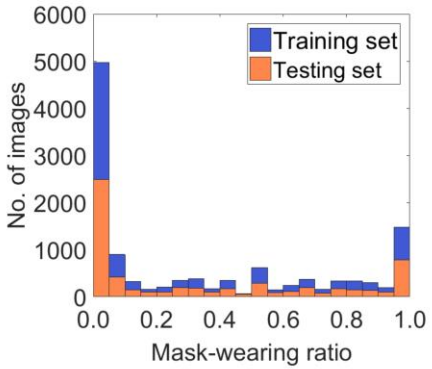
- We created a face mask dataset containing **581,108 face annotations** extracted from **18,088 video frames**.



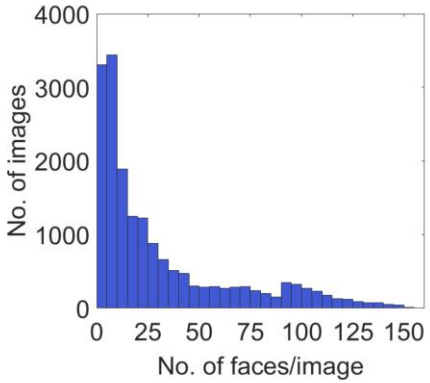
(a) No. of face annotations



(b) Size of face annotations



(c) Mask-wearing ratio



(d) No. of annotated faces per image

Number of annotated faces in NFM dataset.

	Images	Masked	Unmasked	Unknown
Training set	12,058	48,736	317,527	24,594
Testing set	6,030	23,971	154,973	11,307
Total	18,088	72,707	472,500	35,901

Average number of annotated faces per image in NFM dataset

	Masked	Unmasked	Unknown
Training set	4.0	26.3	2.0
Testing set	4.0	25.7	1.9

Large-scale NFM Dataset

- The videos were obtained from the Rambalac YouTube channel: https://www.youtube.com/c/Rambalac/videos
- The videos (approximately 56 hours) were taken in **multiple places**, at **various times**, **before and during the Covid-19 pandemic**.




Evaluation Metrics

- To evaluate prediction accuracy, we used the **Mean Absolute Error (MAE)** and **Pearson correlation** metrics.
  - The MAE metric is defined as $MAE = \frac{1}{N}\sum_{i=1}^{N}\left|c_i - c_i^{gt}\right|$, where N is the total number of images in the testing set, and c_i and c_i^{gt} are the predicted and ground-truth counts, respectively, for image i.
  - The Pearson correlation coefficient γ is defined as

$$\gamma = \frac{\sum_{i=1}^{N}(c_i - \bar{c})(c_i^{gt} - \bar{c}^{gt})}{\sqrt{\sum_{i=1}^{N}(c_i - \bar{c})^2}\sqrt{\sum_{i=1}^{N}(c_i^{gt} - \bar{c}^{gt})^2}}$$

where $\bar{c}$ and $\bar{c}^{gt}$ are the mean of c and c^{gt} respectively.

Likewise, to evaluate the mask-wearing ratio, we computed the Pearson correlation coefficient using the estimated and ground-truth ratios.

Results

- **Comparing the RetinaFace-based method to the CSRNet-based method**
  - The RetinaFace-based method has higher accuracy thanks to its discriminative power (masked faces versus unmasked faces).
  - The CSRNet-based method has a shorter operation time thanks to its compactness.

	RetinaFace			CSRNet		
	MAE	γ	FPS	MAE	γ	FPS
No. of Masked Faces	2.41	**0.81**	0.81	3.42	**0.38**	6.50
No. of Unmasked Faces	10.80	**0.90**	0.81	7.74	**0.92**	6.40
Total No. of Faces	12.55	0.89	0.81	8.46	0.91	6.53
Mask-wearing Ratio	-	**0.94**	**0.81**	-	**0.73**	**3.17**

Results

- **Analyze four scenarios: sparse/dense scenes with low/high ratios:**
  - RetinaFace-based method predicted the accurate ratios for all scenarios.
  - CSRNet suffered in the case of a dense scene with a high ratio.

(a) GT: 0.19, RetinaFace: 0.23, CSRNet: 0.22, faces: 26

(b) GT: 0.06, RetinaFace: 0.04, CSRNet: 0.07, faces: 84

(c) GT: 0.55, RetinaFace: 0.60, CSRNet: 0.51, faces: 11

(d) GT: 0.93, RetinaFace: 0.95, CSRNet: 0.22, faces: 75

Results

- **Location: Shibuya. Recorded time: 10 Mar 2021.**

Conclusion

- We have presented the **first comparative evaluation of detection-based and regression-based methods** for estimating the mask-wearing ratio.
- Evaluation of these methods on **our large-scale face mask dataset (581,108 annotations)** revealed the advantages and disadvantages of each method.
- Future work includes **improving discriminative power** for the regression-based method.

Thank you for your attention!

Dataset is released at: https://zenodo.org/record/5761725