Enhancing Lightweight Neural Networks by Modifying Reference Image Selection for Improved Head Pose Estimation

Nguyen Quang Khai

Instructors

Prof. Wen-Nung Lie

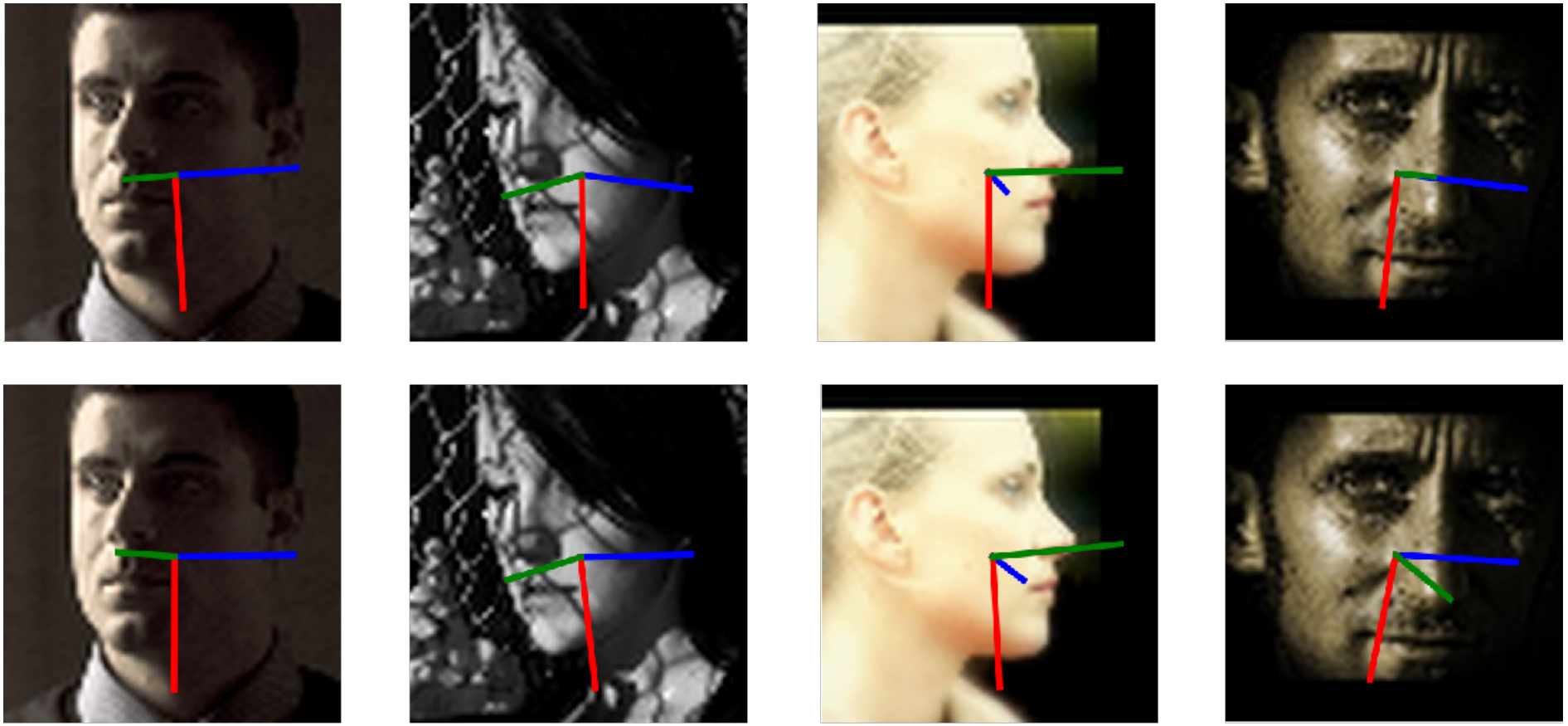
DEPARTMENT OF ELECTRICAL ENGINEERING
NATIONAL CHUNG CHENG UNIVERSITY

7th January 2025

Table of Contents

Table of Contents

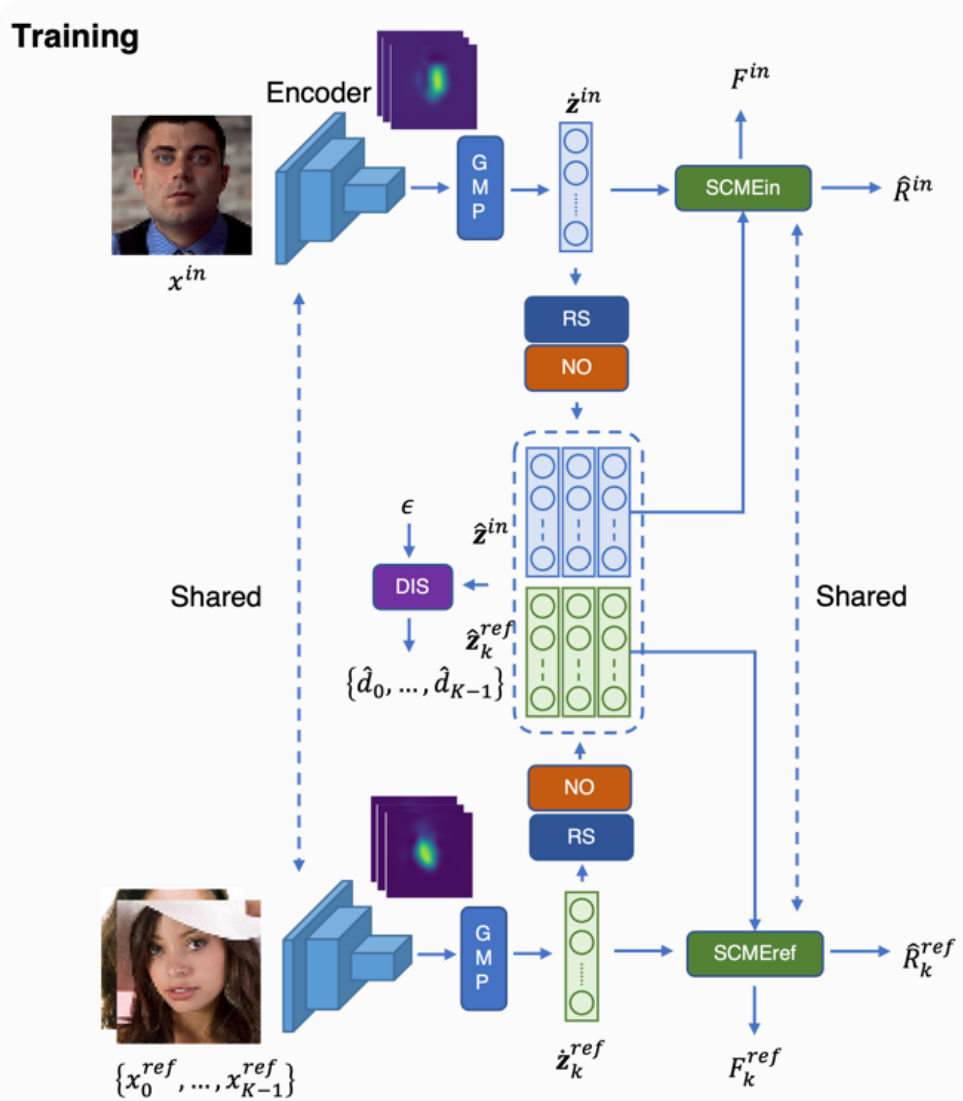
Introduction

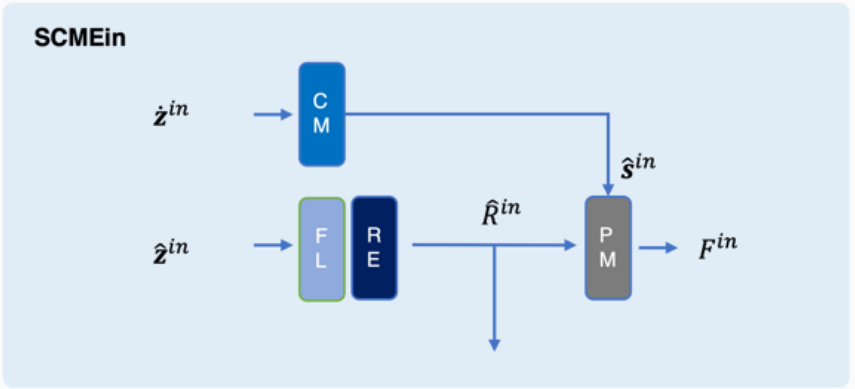


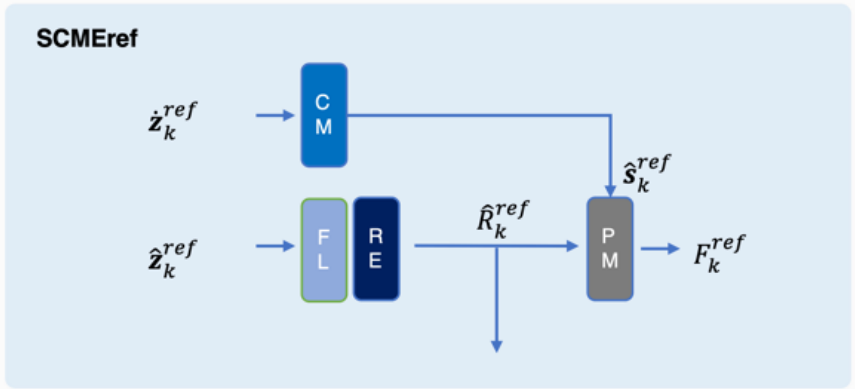
Contributions

- A lightweight neural network is designated for head pose estimation from a single RGB image.
- A novel training strategy based on deep metric learning (DML) is proposed for head pose estimation. Thus, our method is called Metric Head Pose Estimation (MHPE).
- We propose to include a plug-in module, called Separated-Concentration Matrix Fisher distribution Estimation (SCME), to support the backpropagation procedure.
- A new metric called Cost-Error Product (CEP) is proposed to measure the efficiency of a head pose estimation.
- The comprehensive experimental results are demonstrated by following the two common protocols on three public datasets: 300W-LP, AFLW2000, and BIWI.

Training Architecture







Abbreviation

GMP	Global Max Pooling	RE	Rotation Estimator
RS	Reshape	PM	Probability module
NO	Vector Normalization	CM	Concentration module
FL	Flattenning	SCMEin	SCME module for input image
DIS	Distance module	SCMEref	SCME module for references

Distance Module

The triple vector $\mathbf{z}$ is computed by reshaping to a $m \times 3$ matrix and normalizing each column vector. Let $\mathbf{z}^{in}$ and $\mathbf{z}^{ref}$ be the triple-vectors of the input image and the reference image respectively, the distance between these two triple-vectors is defined as:

$$d(\mathbf{z}^{in}, \mathbf{z}^{ref}) = \sqrt{2(3 - \text{trace}(\mathbf{z}^{ref})^T \mathbf{z}^{in})}. \tag{1}$$

where trace$(\cdot)$ is the trace function.
To avoid the problem of class collapse, a uniform distributed noise $\epsilon \sim U[-1, 1]$ is added to the computed distance as expressed in Equation 1.

$$d'(\mathbf{z}^{in}, \mathbf{z}^{ref}) = d(\mathbf{z}^{in}, \mathbf{z}^{ref}) + \epsilon \cdot c \cdot d(\mathbf{z}^{in}, \mathbf{z}^{ref}) \tag{2}$$

where c is a scaling factor.

Inference Architecture

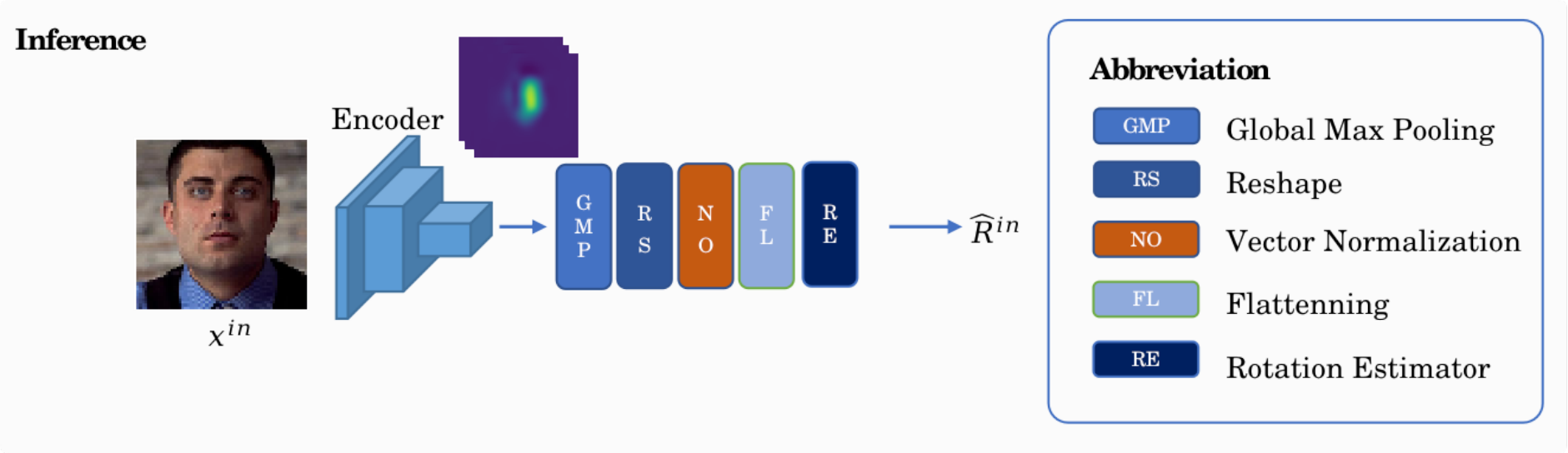


Figure 1.1: Inference phase

- Only the input stream is required to estimate the pose.
- The distance module is ignored in this phase.
- Only the rotation estimator from SCME is used to compute the mean of the predicted rotation matrix.

Encoder

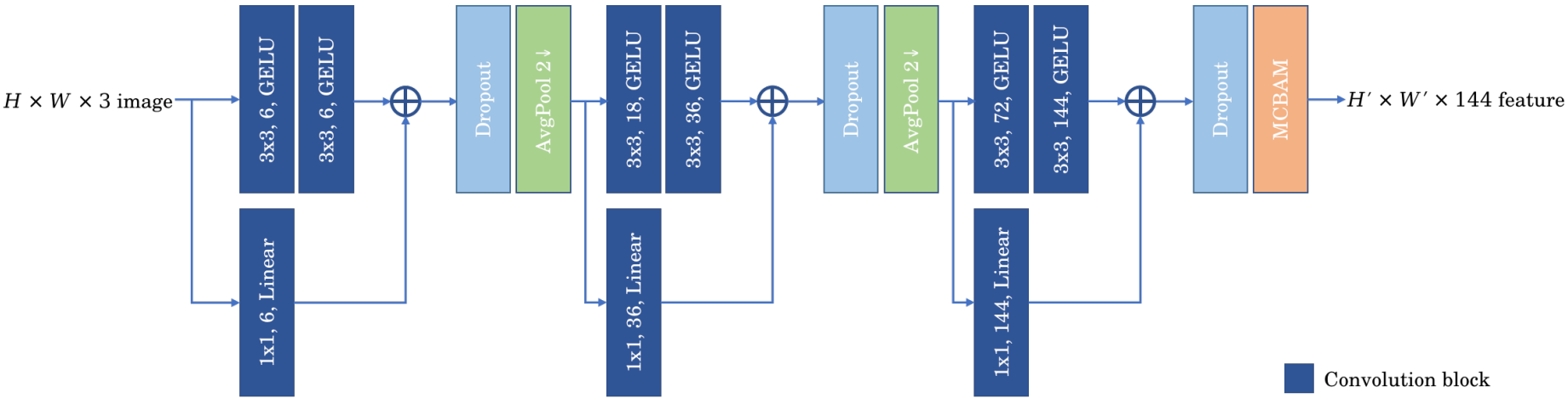


Figure 1.2: The architecture of the encoder.

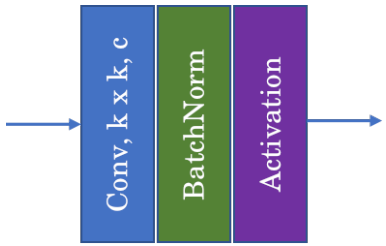


Figure 1.3: Design of a convolution block.

CBAM

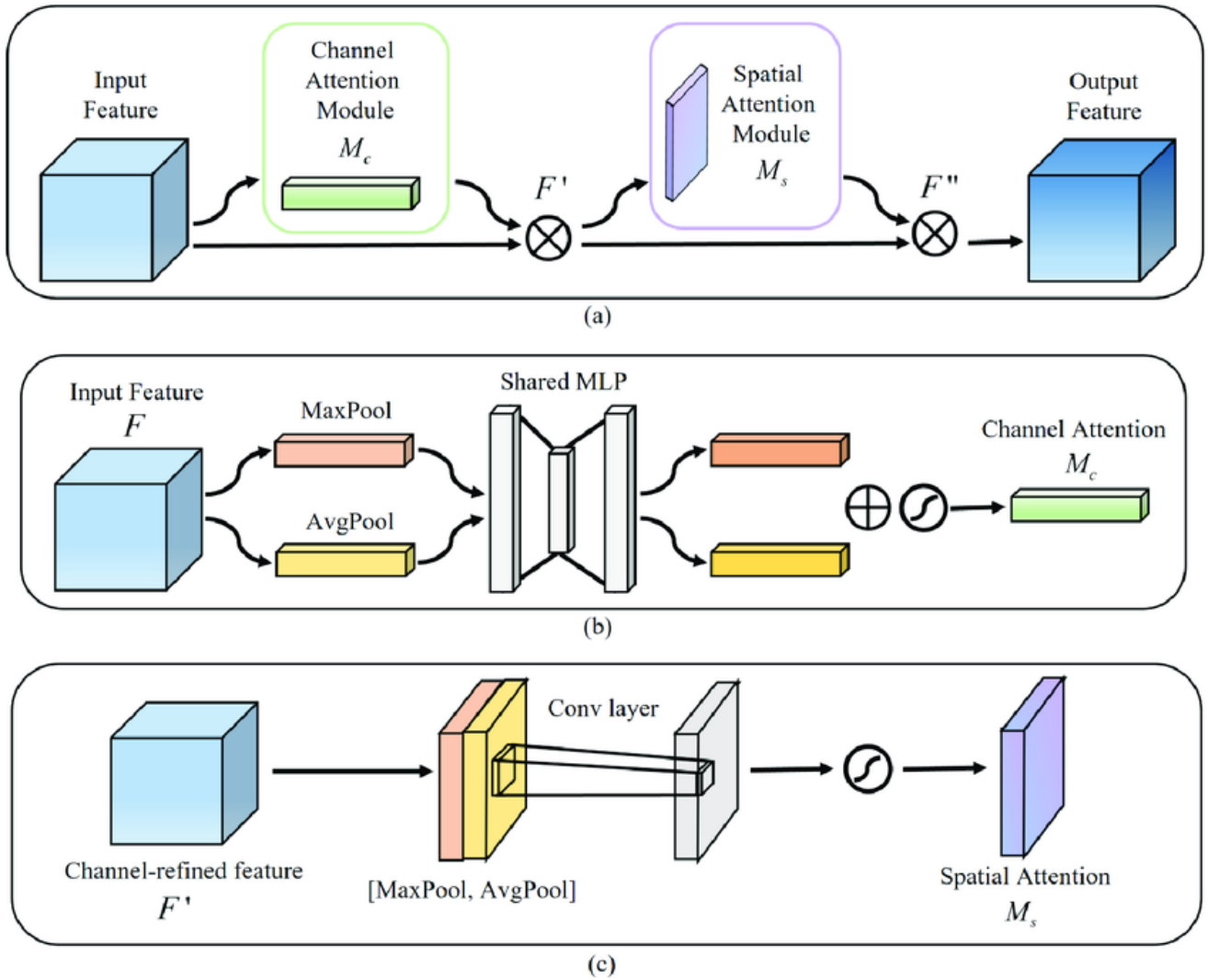


Figure 1.4: The architecture of CBAM.

Modified CBAM (MCBAM)

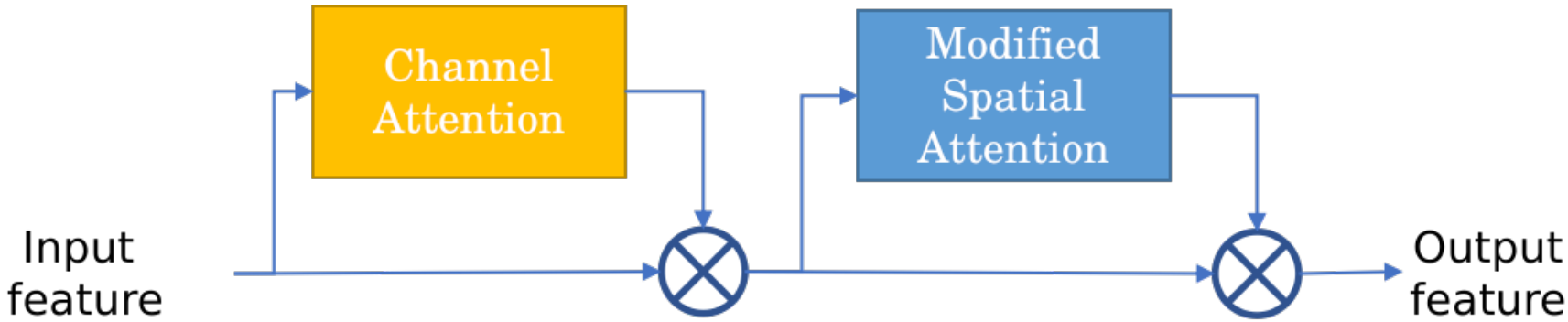


Figure 1.5: The architecture of Modified CBAM.

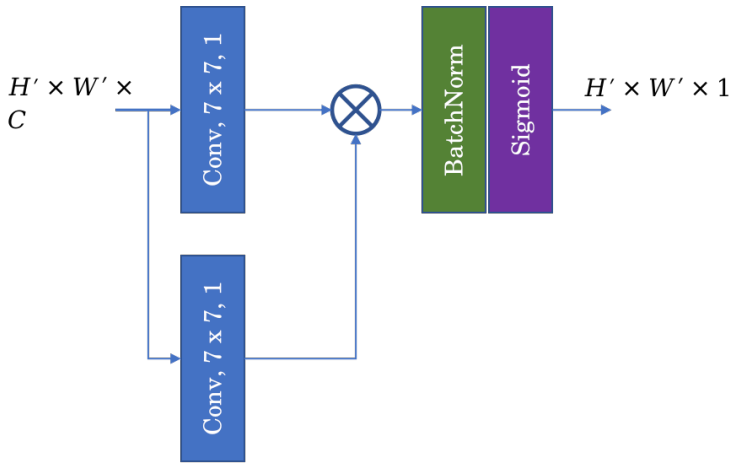


Figure 1.6: Design of Modified Spatial Attention.

Loss Function

Rotation matrix loss

$$\mathcal{L}_{vector} = \frac{1}{3}(\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2^2 + \|\hat{\boldsymbol{v}}_2 - \boldsymbol{v}_2\|_2^2 + \|\hat{\boldsymbol{v}}_3 - \boldsymbol{v}_3\|_2^2) \quad (4)$$

$$\mathcal{L}_{vector\_ortho} = (\boldsymbol{v}_1^T\boldsymbol{v}_2)^2 + (\boldsymbol{v}_2^T\boldsymbol{v}_3)^2 + (\boldsymbol{v}_1^T\boldsymbol{v}_2)^2 + (\|\boldsymbol{v}_1\|_2 - 1)^2 + (\|\boldsymbol{v}_2\|_2 - 1)^2 + (\|\boldsymbol{v}_3\|_2 - 1)^2 \quad (5)$$

where $\hat{R} = [\hat{\boldsymbol{v}}_1 \quad \hat{\boldsymbol{v}}_2 \quad \hat{\boldsymbol{v}}_3]$ is the predicted rotation matrix, $R = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \boldsymbol{v}_3]$ is the ground truth, $\|\,.\,\|_2$ is the L^2 norm.

Feature orthonormality loss

$$\mathcal{L}_{feat\_ortho} = \left\|\hat{\boldsymbol{z}}\hat{\boldsymbol{z}}^{\boldsymbol{T}} - I_3\right\|_F^2 \quad (6)$$

Negative log likelihood loss

$$\mathcal{L}_{nll} = -\log\big(c(F)\big) + trace(F^T R) \quad (7)$$

Loss Function

Distance loss

$$\mathcal{L}_{dist} = \frac{1}{K}\sum_{k=0}^{K-1}\left(\hat{d}\left(\hat{\boldsymbol{z}}^{in}, \hat{\boldsymbol{z}}_k^{ref}\right) - d\left(R^{in}, R_k^{ref}\right)\right)^2 \quad (8)$$

Where R^{in} and R_k^{ref} are the rotation matrix ground truth, $d\left(R^{in}, R_k^{ref}\right) = \sqrt{2(3 - trace\left((R^{in})^T R_k^{ref}\right)}$,

$\hat{d}(.)$ is defined as in Equation 2.

Total loss

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{vector} + \beta\mathcal{L}_{vector\_ortho} + \gamma\,\mathcal{L}_{dist} + \eta\mathcal{L}_{nll} + \mu\mathcal{L}_{feat\_ortho} \quad (9)$$

where α, β, γ, η, μ are the weights for the corresponding sub-losses.

Research Methods

- The primary objective of this research is to enhance the accuracy of head pose estimation (HPE) by using new methods for reference image selection, specifically Gaussian distribution and Uniform distribution, as opposed to the original method of random sampling. By adopting these new reference selection methods, the goal is to reduce the Mean Absolute Error (MAE) and improve the robustness of HPE models across various challenging scenarios.

Table of Contents

Methods

- ***Reference images in bins:*** When selecting reference images, the yaw angle of the input image determines the bin from which reference images are drawn.
- ***Gaussian distribution:*** A method for selecting reference images based on their proximity to a given input yaw angle. This approach aims to give more weight to images whose yaw angles are closer to the input angle, providing a more focused and representative set of reference images.
- ***Reverse Gaussian Distribution:*** This approach aims to give more weight to images whose yaw angles are further to the input angle, reverse with Gaussian distribution.
- ***Uniform Distribution:*** In this method, sample reference images from each bin.

Result in thesis

Hyperparameter	MHPE/MHPE-PB
Batch size	100
Optimizer	Adam
Learning rate	0.01 (multiplied by 0.1 at 20^{th}, 50^{th}, and 80^{th} epoch)
Weight decay	1e-6

Table 1. Hyperparameters in Tai's implementation.

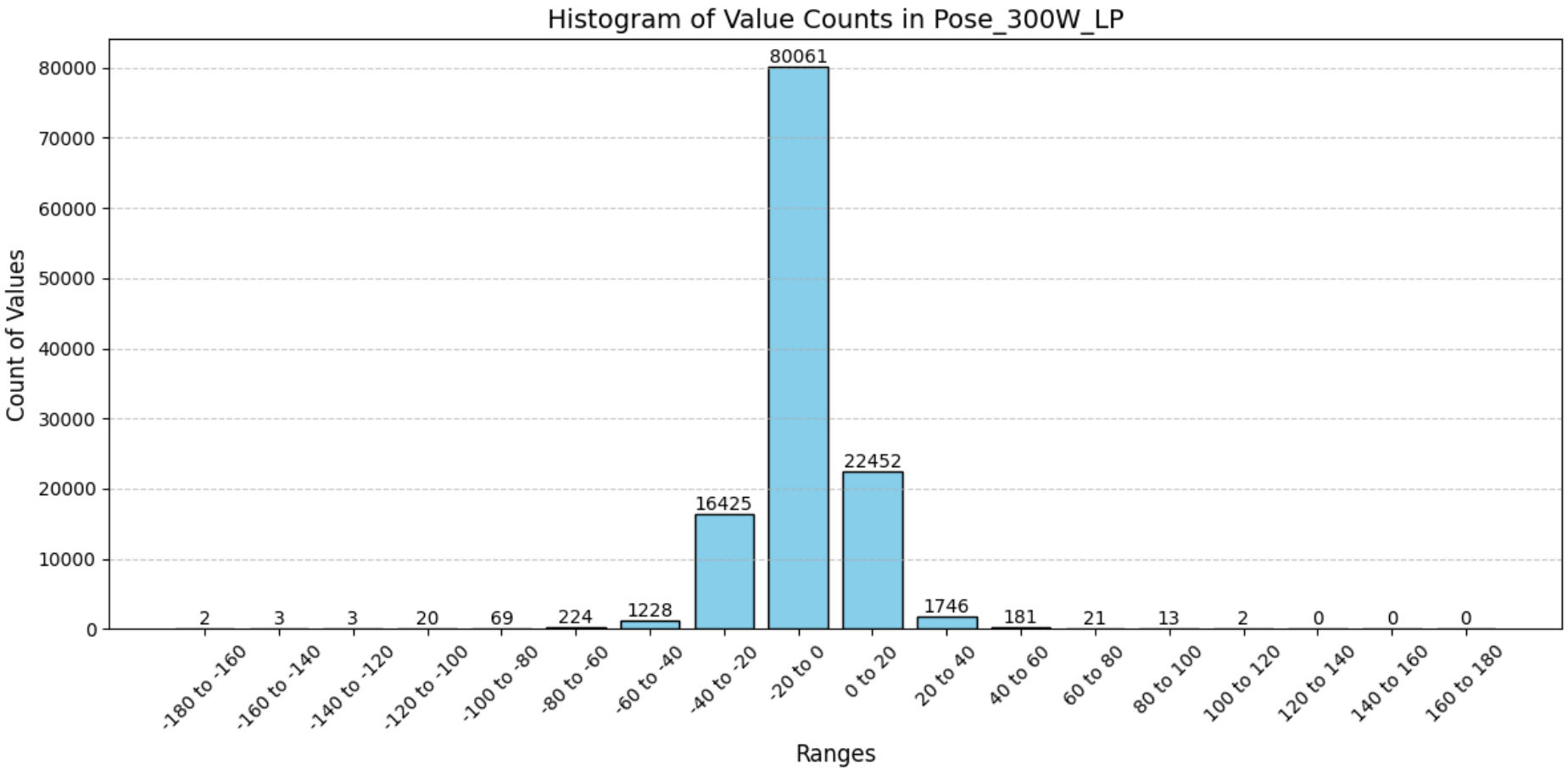
Protocol	Method	Euler angle errors				Vector angle errors			
		Yaw	Pitch	Roll	MAE	Left	Down	Front	MAEV
Protocol 1	**MHPE**	3.99	5.19	3.06	4.08	5.29	5.60	6.76	5.88
Protocol 2	**MHPE**	3.65	5.88	4.23	4.58	5.12	5.29	5.95	5.45

Table 2: MAE research results of Tai

Protocol 1: Training on 300W-LP and testing on BIWI.
Protocol 2: Training on 300W-LP and testing on AFLW2000.

Impact of Using reference images in bins

Using reference images in bins



Using reference images in bins

- **Identify the bin:**
  - Determine the bin corresponding to the yaw angle of the input image.
  - For example, if the input image has yaw $= -10°$, and the bins are divided as $[-20°, 0°]$, $[0°, 20°]$, etc., the image belongs to the bin $[-20°, 0°]$.
- **Retrieve reference images:**
  - Select reference images from the identified bin.
  - If the identified bin does not contain enough reference images, retrieve additional images from the neighboring bins.
  - For example, if the bin $[-20°, 0°]$ is insufficient, additional images can be retrieved from $[0°, 20°]$ or $[-40°, -20°]$.
- This approach ensures that the selected reference images are as relevant as possible while maintaining robustness in cases where certain bins lack sufficient data.

Comparison of MAE Results

Comparison of MAE results between Tai's thesis, Original code and Experiment using reference images in bins.

Result	Range of bins	Protocol	Yaw	Pitch	Roll	MAE
Tai Thesis		Protocol 1	3.99	5.19	3.06	4.08
		Protocol 2	3.65	5.88	4.23	4.58
Original code		Protocol 1	4.01	5.16	3.28	4.13
		Protocol 2	3.85	6.12	4.62	4.71
Experiment	10	Protocol 1	3.64	5.90	3.16	4.23
		Protocol 2	3.76	6.19	4.45	4.83
	20	Protocol 1	3.93	5.22	3.37	4.17
		Protocol 2	3.74	5.97	4.39	4.71
	40	Protocol 1	4.23	4.96	3.20	4.13
		Protocol 2	3.74	6.12	4.41	4.78
	60	Protocol 1	4.06	5.11	3.19	4.11
		Protocol 2	3.79	6.14	4.63	4.81

Impact of Using Gaussian Weights

Using Gaussian Distribution (1/2)

- **Motivation:**
  - In head pose estimation, it is effective to sample reference images closer to the input yaw angle.
  - The Gaussian distribution assigns higher weights to images near the input angle, ensuring relevance and focus.
- **Gaussian Distribution Formula:**

$$W(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where:

- $W(x)$: weight assigned to the yaw angle x,
- μ: mean (input yaw angle),
- σ: standard deviation (controls spread of weights).

Using Gaussian Distribution (2/2)

- **Steps:**
  1. Divide the yaw range $[-180^\circ, 180^\circ]$ into discrete bins.
  2. Compute weights for each bin using the Gaussian formula, with the bin center as x.
  3. Normalize weights: Normalized Weight$(x) = \frac{W(x)}{\sum W(x)}$.
  4. Determine references per bin:
     References from Bin = Normalized Weight$(x) \times k$, where k is the total required references.
  5. Sample references proportional to normalized weights, concentrating around the input angle.

Gaussian Distribution Lookup Table Example

Example: The lookup table divides the yaw angle range $[-180^\circ, 180^\circ]$ into 20° bins. For an input yaw angle $\mu = 10^\circ$, Gaussian weights are computed, and references are sampled accordingly.

Step 1: Setting the Mean (μ) and Standard Deviation (σ):

- $\mu = 10^\circ$ (Gaussian mean), $\sigma = 30^\circ$ (standard deviation).
- Compute weights using: $W(x) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Gaussian Weights for Each Bin

Step 2: Computing Gaussian Weights for Each Bin:

Bin Range	Center (x)	Gaussian Weight
$-20^\circ \to 0^\circ$	-10°	$e^{-\frac{(-10-10)^2}{2(30)^2}} \approx 0.800$
$0^\circ \to 20^\circ$	10°	$e^{-\frac{(10-10)^2}{2(30)^2}} = 1.000$
$20^\circ \to 40^\circ$	30°	$e^{-\frac{(30-10)^2}{2(30)^2}} \approx 0.800$
$40^\circ \to 60^\circ$	50°	$e^{-\frac{(50-10)^2}{2(30)^2}} \approx 0.367$
$-40^\circ \to -20^\circ$	-30°	$e^{-\frac{(-30-10)^2}{2(30)^2}} \approx 0.367$

Normalizing Gaussian Weights

Step 3: Normalizing the Weights:

- Total weight: $0.367 + 0.800 + 1.000 + 0.800 + 0.367 = 3.334$.
- Normalized weights:

Bin Range	Weight	Normalized Weight
$-20^\circ \to 0^\circ$	0.800	$0.800/3.334 \approx 0.240$
$0^\circ \to 20^\circ$	1.000	$1.000/3.334 \approx 0.300$
$20^\circ \to 40^\circ$	0.800	$0.800/3.334 \approx 0.240$
$40^\circ \to 60^\circ$	0.367	$0.367/3.334 \approx 0.110$
$-40^\circ \to -20^\circ$	0.367	$0.367/3.334 \approx 0.110$

Sampling References

Step 4: Sampling References:

- For $k = 10$ references, the number of images sampled per bin is:

$$\text{Images from Bin} = \text{Normalized Weight} \times k.$$

- Distribution:

Bin Range	**Normalized Weight**	**Images to Sample ($k = 10$)**
$-20^\circ \to 0^\circ$	0.240	$0.240 \times 10 \approx 2$
$0^\circ \to 20^\circ$	0.300	$0.300 \times 10 \approx 3$
$20^\circ \to 40^\circ$	0.240	$0.240 \times 10 \approx 2$
$40^\circ \to 60^\circ$	0.110	$0.110 \times 10 \approx 1$
$-40^\circ \to -20^\circ$	0.110	$0.110 \times 10 \approx 1$

Final Sampling

Step 5: Final Sampling:

- 3 references from $0^\circ \rightarrow 20^\circ$,
- 2 references each from $-20^\circ \rightarrow 0^\circ$ and $20^\circ \rightarrow 40^\circ$,
- 1 reference each from $40^\circ \rightarrow 60^\circ$ and $-40^\circ \rightarrow -20^\circ$.

Comparison of MAE Results by Gaussian Distribution

Comparison of MAE results between Tai's thesis, Original code and Experiment using reference images by Gaussian distribution.

Result	Range of bins	Std Dev	Protocol	Yaw	Pitch	Roll	MAE
Tai Thesis			Protocol 1	3.99	5.19	3.06	4.08
			Protocol 2	3.65	5.88	4.23	4.58
Original code			Protocol 1	4.01	5.16	3.28	4.13
			Protocol 2	3.85	6.12	4.62	4.71
Experiment	20	15	Protocol 1	3.72	5.41	3.30	4.14
			Protocol 2	4.09	6.30	4.60	4.95
	20	20	Protocol 1	3.65	5.49	3.14	4.09
			Protocol 2	3.91	6.10	4.58	4.71
	20	30	Protocol 1	3.58	5.30	3.10	3.98
			Protocol 2	3.91	6.09	4.52	4.69
	20	37.5	Protocol 1	3.60	5.39	3.11	4.03
			Protocol 2	3.90	6.09	4.58	4.78
	20	45	Protocol 1	3.83	5.39	3.34	4.22
			Protocol 2	3.83	6.36	4.61	4.92

Impact of Using Inverse Gaussian Weights

Inverse Gaussian Weights

- In the inverse Gaussian method, weights are computed such that the weight distribution is the opposite of the standard Gaussian distribution.
- Bins farther from the mean value (μ) have higher weights, while bins closer to the mean have lower weights.
- Given yaw bins with a mean (μ) of 10° and a standard deviation (σ) of 30°, the weights for each bin are computed using the inverse Gaussian formula:

$$W(x) = e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Yaw Bin Weights

- Below is the weight table for the yaw bins:

Bin Range	Weight (Inverse Gaussian)
$-20^\circ \rightarrow 0^\circ$	1.56
$0^\circ \rightarrow 20^\circ$	1.00
$20^\circ \rightarrow 40^\circ$	1.56
$40^\circ \rightarrow 60^\circ$	5.93
$-40^\circ \rightarrow -20^\circ$	5.93

Normalized Weight Table

- After normalizing the weights to ensure their sum equals 1, we get the following normalized weight table:
- Total weight $= 1.56 + 1.00 + 1.56 + 5.93 + 5.93 = 16.98$

Bin Range	Normalized Weight	Number of References ($k = 10$)
$-20^\circ \rightarrow 0^\circ$	0.092	1
$0^\circ \rightarrow 20^\circ$	0.059	1
$20^\circ \rightarrow 40^\circ$	0.092	1
$40^\circ \rightarrow 60^\circ$	0.349	3
$-40^\circ \rightarrow -20^\circ$	0.349	3

- References are sampled more heavily from bins farther from the mean value (μ) and less from bins closer to it.

Comparison of MAE Results by Reverse Gaussian Distribution

Comparison of MAE results between Tai's thesis, Original code and Experiment using reference images by Reverse Gaussian distribution.

Result	Range of bins	Std Dev	Protocol	Yaw	Pitch	Roll	MAE
Tai Thesis			Protocol 1	3.99	5.19	3.06	4.08
			Protocol 2	3.65	5.88	4.23	4.58
Original code			Protocol 1	4.01	5.16	3.28	4.13
			Protocol 2	3.85	6.12	4.62	4.71
	20	20	Protocol 1	9.85	7.05	4.48	7.13
			Protocol 2	11.45	8.85	7.39	9.24
	20	30	Protocol 1	6.29	10.55	4.44	7.09
			Protocol 2	10.09	8.71	7.46	8.75
	20	45	Protocol 1	5.37	8.16	4.56	6.03
			Protocol 2	6.23	8.01	6.62	6.96

Impact of Using Uniform Distribution

Using Uniform Distribution

- In this method, we use a uniform distribution to sample reference images from each bin.
- Unlike the Gaussian distribution where weights are calculated based on distance from the mean, here we simply select one image from each bin until the required number of references (k) is reached.
- If there are fewer bins than the required references, the process repeats, sampling from the bins again until the total number of references is collected.

Steps for Uniform Distribution Sampling

- **Step 1:** Divide the yaw angle range $[-180°, 180°]$ into equal-sized bins.
- **Step 2:** From each bin, one reference image is selected. This means that for each bin, you choose one image that corresponds to the yaw angle within the range of that bin.
- **Step 3:** If the number of bins is insufficient to meet the required references (k), the process repeats, sampling from the bins in the same order until k references are collected.

Example of Uniform Distribution Sampling

- **Bins:** Bin 1, Bin 2, Bin 3, Bin 4, Bin 5
- **References Sampled:**
  1. Bin 1 $\rightarrow$ Reference 1
  2. Bin 2 $\rightarrow$ Reference 2
  3. Bin 3 $\rightarrow$ Reference 3
  4. Bin 4 $\rightarrow$ Reference 4
  5. Bin 5 $\rightarrow$ Reference 5
  6. Bin 1 $\rightarrow$ Reference 6
  7. Bin 2 $\rightarrow$ Reference 7
  8. Bin 3 $\rightarrow$ Reference 8
  9. Bin 4 $\rightarrow$ Reference 9
  10. Bin 5 $\rightarrow$ Reference 10

Comparison of MAE Results by Uniform Distribution

Comparison of MAE results between Tai's thesis, Original code and Experiment using reference images in bins.

Result	Range of bins	Protocol	Yaw	Pitch	Roll	MAE
Tai Thesis		Protocol 1	3.99	5.19	3.06	4.08
		Protocol 2	3.65	5.88	4.23	4.58
Original code		Protocol 1	4.01	5.16	3.28	4.13
		Protocol 2	3.85	6.12	4.62	4.71
	20	Protocol 1	18.57	14.67	7.57	13.61
		Protocol 2	18.91	13.62	9.54	14.03

Table of Contents

Conclusion Results

Result	Range of bins	Std Dev	Protocol	Yaw	Pitch	Roll	MAE
Tai Thesis	-	-	Protocol 1	3.99	5.19	3.06	4.08
			Protocol 2	3.65	5.88	4.23	4.58
Original code	-	-	Protocol 1	4.01	5.16	3.28	4.13
			Protocol 2	3.85	6.12	4.62	4.71
Reference images in bins	20	-	Protocol 1	4.21	5.13	3.16	4.17
			Protocol 2	3.61	6.07	4.34	4.71
Gaussian distribution	20	30	Protocol 1	3.58	5.30	3.10	3.98
			Protocol 2	3.91	6.09	4.58	4.69
Reverse Gaussian distribution	20	45	Protocol 1	5.37	8.16	4.56	6.03
			Protocol 2	6.23	8.01	6.62	6.69
Uniform distribution	20	-	Protocol 1	18.57	14.67	7.57	13.61
			Protocol 2	18.91	16.62	9.54	14.03

Conclusion

- Replacing random sampling with Gaussian distribution showed significant improvement in results.
- Optimal performance was achieved with a standard deviation of 30 across both protocols.
- The inverse Gaussian distribution led to worse results, indicating the importance of selecting reference images based on proximity to the input yaw angle.
- Uniform distribution, where one image is selected per bin, resulted in inefficient learning due to lack of focus on critical yaw angles.
- Grouping reference images into bins based on proximity to the input yaw angle showed similar accuracy to the original results, but lacked diversity.
- A combination of both similar and diverse features in reference images leads to optimal model performance.

Thank you for your attention!