8. Supplementary

In the main paper, the discussion on conventional open-set backdoor performance was restricted to the mask trigger on the LFW and WebFace datasets due to space constraints. However, it was noted that similar results were observed for the other triggers, offering only marginal added value to the main document as the resulting statements remain the same while adding several extra pages.

To provide the reader with a comprehensive understanding of the results for all conventional backdoor approaches in an open-set setting for each trigger, we will now present the remaining results. This further underscores our assertion that current state-of-the-art backdoor attacks designed for closed-set classification scenarios do not translate effectively to real-world open-set recognition tasks, reinforcing the significance of our proposed contribution.

Tables 7, 8, 9, and 10 show the results on LFW and Tables 11, 12, 13, and 14 show the results for digital sunglasses, hat, physical sunglasses, and red square triggers. There, the open-set performances of conventional backdoor attacks are shown for the four introduced face recognition models. Value N still represents the number of trained identities to distinguish from and the error on clean data represents how well the model can recognize people without the presence of a trigger and the error on backdoor data shows how well the model can link the trigger to the target identity. Performance is shown in terms of FNMR at different FMRs.

In all cases, it can be seen that training backdoor attacks on the classification model leads to unreasonably high errors on clean data, i.e. the resulting face recognition system does not perform well. Compared the performance of the original face recognition system (Table 5 in the main paper), it is unlike that these systems would be used in a real-world context due to their low performance. Moreover, the performance of the backdoor itself is low as and only leads to low error rates for high FMRs. These results motivate the need for more effective open-set backdoor attacks in more realistic scenarios to develop more effective defense mechanisms for real-world applications. These results reflect what we have already found out in the main section and are only included here for the sake of completeness.

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	96.0	91.1	77.6	44.9	84.0	92.9	83.3	65.7	30.3	90.7
	10	93.5	82.9	60.7	25.0	90.8	91.5	78.6	54.4	17.2	93.3
	25	79.9	62.1	34.7	12.1	94.8	93.2	75.5	37.6	7.8	96.3
	50	79.9	60.7	34.6	11.6	95.6	93.5	78.5	39.8	7.9	95.6
	75	71.2	48.5	25.0	7.9	96.2	95.8	80.5	40.0	9.2	95.2
	100	77.1	57.5	32.1	12.9	94.9	94.4	81.1	39.9	10.6	94.2
ArcFace	5	99.0	97.5	89.1	56.3	76.7	81.6	72.4	60.5	26.5	92.6
	10	96.2	90.2	73.5	38.1	84.7	77.2	66.0	38.0	6.7	96.5
	25	85.1	67.2	44.0	16.3	93.7	73.5	47.2	18.0	4.8	97.2
	50	81.7	68.4	47.4	21.2	91.5	52.4	21.9	6.4	4.4	97.7
	75	64.7	42.9	23.3	7.9	97.0	70.9	35.6	10.9	6.8	96.6
	100	77.6	59.1	33.4	10.7	95.4	65.9	29.5	10.1	2.5	98.7
MagFace	5	99.0	96.3	88.2	61.7	73.9	91.6	83.3	74.6	65.1	69.8
	10	96.0	87.6	68.6	33.2	87.5	93.3	86.4	79.4	60.4	79.0
	25	85.7	73.3	51.8	22.2	91.6	92.2	88.9	74.6	34.5	88.3
	50	90.7	78.3	54.0	19.7	92.6	95.8	91.7	74.9	22.0	92.1
	75	91.6	79.0	58.0	24.0	91.0	96.8	91.0	72.4	32.5	88.7
	100	90.8	77.3	53.0	20.8	92.6	97.5	93.1	76.2	34.7	88.0
QMagFace	5	99.8	98.3	93.1	62.5	75.9	99.6	97.4	85.5	68.5	70.0
	10	99.4	93.5	78.6	38.5	85.6	99.7	96.6	81.2	61.1	77.3
	25	95.5	87.9	68.6	31.5	88.0	96.4	92.3	80.0	33.5	89.0
	50	96.1	83.3	57.3	23.4	91.8	96.8	92.9	77.3	35.4	87.5
	75	95.7	87.7	65.4	28.9	89.6	97.6	92.8	71.0	20.4	92.2
	100	92.1	75.6	47.1	17.0	93.5	96.9	92.7	74.2	26.3	89.8

Table 7. Conventional Open-Set Backdoor Performance on LFW for the digital sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the digital sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		LFW										WebFace									
		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data					Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	97.1	92.7	80.3	47.4	82.1	98.5	89.5	65.9	15.6	95.1	97.7	94.0	83.7	53.9	78.1	97.8	88.5	61.8	11.4	95.7
	10	95.0	83.7	60.8	25.0	90.9	98.1	90.2	46.8	4.0	97.8	95.4	89.0	71.7	39.7	84.7	97.6	85.4	43.6	9.3	96.9
	25	77.4	59.8	33.9	10.6	96.0	95.1	80.4	33.6	3.0	98.3	88.3	75.5	54.8	25.8	90.3	96.8	79.7	44.5	9.5	96.8
	50	66.8	49.4	27.7	8.0	97.1	96.5	72.9	33.4	5.4	97.8	83.6	69.9	49.0	22.9	91.1	93.2	78.7	41.6	8.2	97.0
	75	64.8	42.6	20.2	4.5	98.1	96.4	70.7	23.9	2.0	98.9	83.8	68.4	46.8	21.7	91.3	96.4	78.4	44.9	11.6	96.0
	100	67.3	44.6	21.9	5.5	97.9	92.4	71.6	21.9	1.3	99.0	83.9	68.2	47.3	22.3	91.1	92.3	80.3	47.4	11.7	95.9
ArcFace	5	99.3	97.9	90.6	58.5	74.1	93.2	72.8	25.2	0.1	99.2	99.1	96.5	87.8	56.9	78.0	89.3	71.4	33.3	1.5	98.6
	10	95.9	86.5	65.5	28.2	90.0	89.2	47.9	2.9	0.0	99.8	96.3	89.6	71.9	36.5	86.4	72.2	31.7	3.6	0.0	99.8
	25	83.8	66.7	41.9	11.6	95.4	52.6	10.2	0.0	0.0	100.0	91.6	80.2	58.7	27.5	89.2	78.4	32.3	6.4	0.3	99.7
	50	61.9	44.5	21.8	4.7	98.3	59.9	17.4	0.2	0.0	99.9	79.7	65.2	44.6	20.3	91.7	54.2	26.0	5.0	0.0	99.8
	75	61.2	39.6	20.8	6.4	97.7	42.6	14.2	1.2	0.0	99.9	72.7	57.4	38.3	17.9	92.7	70.6	26.2	7.6	0.9	99.5
	100	57.4	41.0	19.7	5.2	98.3	49.0	26.4	4.0	0.0	99.8	68.6	52.3	33.9	15.7	93.5	58.4	27.0	5.8	0.3	99.7
MagFace	5	99.4	97.8	92.3	66.3	72.1	99.8	98.8	93.4	59.6	80.2	99.3	97.5	91.9	68.0	70.0	99.5	97.2	87.1	47.2	85.2
	10	98.8	95.9	84.4	54.0	79.5	98.6	94.7	77.2	31.0	91.6	97.2	92.4	78.8	45.9	82.3	98.2	87.7	54.2	13.4	95.1
	25	88.2	70.8	44.9	14.2	94.3	96.4	87.6	55.6	12.8	95.7	85.6	70.3	49.3	23.8	90.4	89.7	78.0	44.1	11.1	95.4
	50	76.5	56.7	30.3	7.9	97.0	97.5	77.2	38.1	5.2	97.9	76.1	58.3	38.6	17.2	92.7	93.4	72.0	38.8	8.4	96.9
	75	79.8	62.5	35.5	10.0	96.1	96.3	84.5	56.1	15.1	95.1	70.5	56.4	37.6	19.1	91.6	94.2	64.7	29.2	7.5	97.5
	100	75.4	53.8	32.1	10.6	95.4	96.4	86.3	51.5	9.5	96.4	73.9	56.4	38.0	18.4	92.2	90.7	75.5	39.7	8.8	96.6
QMagFace	5	99.9	98.6	89.3	59.6	77.1	97.6	92.5	79.7	62.3	80.0	99.8	98.2	90.2	58.9	77.1	98.8	97.4	91.5	61.4	79.1
	10	99.0	93.2	77.2	41.9	84.2	97.5	85.3	72.9	45.1	81.3	99.2	92.4	73.4	37.7	85.2	95.5	87.2	65.5	25.4	92.4
	25	88.5	68.7	43.4	16.5	93.5	89.0	69.7	34.9	5.2	97.8	92.0	78.0	55.2	26.4	89.6	75.4	48.7	12.1	0.1	99.6
	50	88.2	71.4	43.4	17.7	93.2	90.0	69.1	19.4	0.8	99.2	84.2	68.4	48.1	22.5	90.4	71.3	38.5	8.9	0.3	99.6
	75	86.5	65.5	39.5	15.1	94.1	92.8	75.4	30.4	2.2	98.6	87.1	73.5	51.9	24.3	90.0	73.0	39.4	9.7	0.9	99.5
	100	85.7	63.3	37.0	13.0	94.5	90.0	60.7	13.0	0.2	99.5	85.6	71.1	49.1	23.3	90.4	89.3	52.7	14.1	0.6	99.4

Table 8. Conventional Open-Set Backdoor Performance on LFW and WebFace for the hat trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the hat trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	25	96.6	92.4	80.8	50.0	81.3	99.1	91.9	61.6	16.2	94.8
	50	94.5	84.4	58.7	24.2	91.9	96.3	85.3	38.8	3.0	98.3
	75	82.4	65.9	37.6	11.2	96.0	91.0	69.5	24.3	1.3	99.0
	100	69.8	51.8	27.0	6.6	97.4	94.7	73.6	15.3	0.3	99.4
	5	62.7	40.6	19.1	4.6	98.2	90.6	55.8	15.3	0.4	99.4
	25	67.8	46.0	22.7	7.1	97.6	90.5	63.0	12.3	0.2	99.5
ArcFace	5	99.0	98.1	90.1	59.5	74.9	90.4	73.7	35.0	0.7	98.7
	25	96.1	87.9	70.2	31.6	88.5	76.6	47.6	10.9	0.0	99.6
	50	74.7	58.0	34.5	7.8	96.8	39.9	1.6	0.0	0.0	100.0
	75	67.9	47.6	25.6	6.5	97.5	27.8	7.5	0.3	0.0	100.0
	100	59.0	35.6	15.3	3.5	98.6	39.7	18.4	3.8	0.2	99.8
	5	56.1	34.9	16.9	4.3	98.3	45.0	17.6	3.7	0.2	99.8
MagFace	50	99.2	97.3	91.0	67.6	72.3	99.2	97.5	89.1	51.1	85.4
	75	99.0	95.3	85.9	47.2	84.1	99.4	96.4	82.8	37.4	89.8
	100	84.8	69.2	41.0	13.1	95.1	92.9	80.0	53.6	11.9	95.9
	5	76.1	58.8	32.4	10.0	95.8	92.9	82.8	51.4	10.9	96.3
	25	76.1	53.1	28.0	9.1	96.8	95.1	85.8	58.8	12.4	95.8
	50	65.9	45.6	22.7	6.9	97.1	92.0	81.3	52.5	9.7	96.5
QMagFace	75	100.0	99.3	96.1	78.5	66.1	100.0	99.8	97.3	72.5	76.2
	100	99.5	98.0	87.9	56.2	79.0	99.2	96.2	86.0	40.6	89.0
	5	92.0	75.2	48.5	16.6	94.0	97.9	87.2	49.5	6.3	97.3
	25	75.8	55.2	30.0	9.1	96.2	93.6	76.1	32.5	3.0	98.4
	50	82.3	62.0	35.2	11.3	96.0	92.6	84.6	54.8	9.7	96.5
	75	76.7	56.9	32.7	11.5	95.6	96.4	85.0	51.7	11.2	95.7

Table 9. Conventional Open-Set Backdoor Performance on LFW for the physical sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the physical sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	25	97.0	91.9	78.2	45.0	82.9	96.2	90.2	78.9	55.4	76.4
	50	92.9	83.6	65.5	30.0	88.5	95.7	89.7	78.3	44.7	85.6
	75	80.1	63.1	39.2	14.4	94.4	94.1	85.9	59.7	18.5	94.0
	100	78.2	58.9	33.5	10.5	95.8	98.3	90.2	64.0	18.9	93.7
	5	81.4	63.8	37.9	12.1	95.0	98.0	91.1	69.5	22.2	92.6
	25	88.4	73.2	48.2	17.6	93.4	98.9	94.0	78.0	33.2	89.0
ArcFace	5	98.6	96.3	88.1	56.8	77.0	85.5	75.6	69.5	58.8	75.5
	25	98.0	92.7	76.7	39.5	85.9	81.8	79.2	73.8	29.6	92.2
	50	93.5	83.0	64.5	31.3	88.3	87.3	75.0	31.7	0.4	99.0
	75	89.0	77.2	55.6	22.2	91.8	73.1	41.2	7.7	0.4	99.6
	100	83.1	69.2	47.5	15.6	94.5	63.8	30.2	7.5	0.4	99.7
	5	81.7	66.0	38.4	12.3	95.4	53.3	21.2	3.7	1.0	99.6
MagFace	50	98.9	96.3	87.1	57.6	77.1	90.5	81.2	73.0	70.0	51.9
	75	97.4	91.2	74.4	41.4	85.1	86.5	82.0	79.7	77.8	49.7
	100	90.8	79.1	57.7	21.7	91.7	91.7	90.3	89.1	85.1	48.8
	5	91.9	80.6	55.8	23.6	91.7	95.2	94.2	91.6	76.5	58.0
	25	91.0	79.5	57.0	23.7	91.4	97.4	96.3	93.6	78.8	53.4
	50	91.3	79.8	58.9	24.9	91.5	97.7	96.0	91.5	77.7	56.6
QMagFace	75	99.4	97.9	91.0	57.4	78.4	99.8	95.1	82.9	70.8	51.1
	100	96.7	91.3	77.1	45.2	83.7	98.0	86.0	80.3	78.1	46.2
	5	90.0	72.7	45.3	16.4	94.3	99.3	90.9	87.2	81.4	52.0
	25	96.0	83.4	56.0	17.4	93.5	96.8	95.1	93.1	82.2	52.1
	50	98.8	86.8	55.9	18.3	93.0	97.7	96.3	93.0	79.0	55.5
	75	98.7	91.3	74.0	31.2	89.0	99.4	98.3	95.1	83.0	51.7

Table 10. Conventional Open-Set Backdoor Performance on LFW for the red square trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the LFW dataset for the red square trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	98.1	94.6	84.3	54.2	78.8	97.5	92.2	79.7	40.5	87.9
	10	97.7	93.7	79.8	48.6	79.6	96.9	89.9	60.9	8.7	95.7
	25	90.3	79.0	60.4	32.0	86.4	96.8	89.0	58.0	7.9	96.5
	50	87.9	75.5	54.6	26.1	89.4	97.6	87.1	53.2	8.1	96.2
	75	87.5	74.7	55.4	29.5	87.8	97.8	87.1	49.3	4.1	97.5
	100	89.4	76.8	57.1	29.0	87.6	98.5	90.0	49.4	3.6	97.7
ArcFace	5	98.9	96.4	88.5	60.3	76.0	90.9	85.3	73.7	33.3	90.5
	10	96.8	92.0	78.7	46.6	81.5	88.6	73.9	27.4	2.6	98.1
	25	95.0	87.0	72.1	41.3	82.6	88.4	48.8	7.8	3.3	98.3
	50	83.6	71.1	51.3	24.7	89.9	44.4	18.6	3.2	1.2	99.3
	75	76.4	62.8	43.9	21.0	91.2	44.4	16.9	6.0	3.7	98.2
	100	70.8	57.2	39.1	18.3	92.2	57.3	22.6	5.3	1.9	98.9
MagFace	5	98.7	95.1	84.6	55.2	77.4	93.6	88.5	82.5	70.6	67.0
	10	97.2	92.3	79.3	47.5	80.9	96.0	92.0	81.5	41.0	88.7
	25	93.2	85.1	67.7	35.7	85.2	98.1	95.4	82.9	41.4	86.8
	50	95.4	86.8	67.4	34.5	85.2	98.8	95.1	79.1	30.8	90.1
	75	93.4	84.6	66.3	35.7	85.3	98.4	93.3	74.8	28.7	89.5
	100	92.4	83.4	65.3	32.9	85.9	98.7	94.6	76.8	25.6	92.0
QMagFace	5	99.8	97.2	86.6	56.4	76.9	97.9	94.1	83.5	60.5	80.8
	10	99.1	94.9	79.7	44.0	82.1	98.3	95.2	89.0	63.1	80.9
	25	98.7	91.7	70.5	36.4	84.6	98.9	94.8	79.1	27.5	91.1
	50	96.4	86.4	63.4	32.6	86.1	98.2	92.4	69.5	20.7	92.3
	75	97.3	90.0	71.0	36.8	84.7	99.4	95.6	77.3	29.2	90.5
	100	97.1	91.1	73.5	39.0	83.7	98.7	93.3	72.5	25.5	91.6

Table 11. Conventional Open-Set Backdoor Performance on WebFace for the digitial sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the digital sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	97.7	94.0	83.7	53.9	78.1	97.8	88.5	61.8	11.4	95.7
	10	95.4	89.0	71.7	39.7	84.7	97.6	85.4	43.6	9.3	96.9
	25	88.3	75.5	54.8	25.8	90.3	96.8	79.7	44.5	9.5	96.8
	50	83.6	69.9	49.0	22.9	91.1	93.2	78.7	41.6	8.2	97.0
	75	83.8	68.4	46.8	21.7	91.3	96.4	78.4	44.9	11.6	96.0
	100	83.9	68.2	47.3	22.3	91.1	92.3	80.3	47.4	11.7	95.9
ArcFace	5	99.1	96.5	87.8	56.9	78.0	89.3	71.4	33.3	1.5	98.6
	10	96.3	89.6	71.9	36.5	86.4	72.2	31.7	3.6	0.0	99.8
	25	91.6	80.2	58.7	27.5	89.2	78.4	32.3	6.4	0.3	99.7
	50	79.7	65.2	44.6	20.3	91.7	54.2	26.0	5.0	0.0	99.8
	75	72.7	57.4	38.3	17.9	92.7	70.6	26.2	7.6	0.9	99.5
	100	68.6	52.3	33.9	15.7	93.5	58.4	27.0	5.8	0.3	99.7
MagFace	5	99.3	97.5	91.9	68.0	70.0	99.5	97.2	87.1	47.2	85.2
	10	97.2	92.4	78.8	45.9	82.3	98.2	87.7	54.2	13.4	95.1
	25	85.6	70.3	49.3	23.8	90.4	89.7	78.0	44.1	11.1	95.4
	50	76.1	58.3	38.6	17.2	92.7	93.4	72.0	38.8	8.4	96.9
	75	70.5	56.4	37.6	19.1	91.6	94.2	64.7	29.2	7.5	97.5
	100	73.9	56.4	38.0	18.4	92.2	90.7	75.5	39.7	8.8	96.6
QMagFace	5	99.9	99.5	94.9	69.9	70.1	99.9	99.8	95.8	66.2	79.4
	10	97.0	91.4	75.0	40.9	83.5	98.9	91.9	64.5	18.2	94.1
	25	84.2	70.5	49.7	23.4	90.4	93.7	75.6	40.1	8.3	97.0
	50	76.0	60.4	40.4	20.4	90.9	95.6	76.8	42.5	7.7	96.9
	75	75.0	58.7	37.9	17.9	92.4	91.2	71.6	39.9	11.2	95.6
	100	73.7	56.1	36.7	16.6	92.8	91.4	74.8	40.3	8.2	97.1

Table 12. Conventional Open-Set Backdoor Performance on WebFace for the hat trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the hat trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	98.3	95.3	84.7	56.5	77.8	99.3	77.6	58.4	7.5	96.9
	10	96.7	90.8	73.9	39.6	84.6	90.2	62.5	65.9	19.0	93.7
	25	89.9	78.2	58.1	28.5	88.4	87.4	46.2	53.2	10.6	96.3
	50	84.5	71.1	50.7	22.6	90.9	84.7	58.0	50.1	6.5	97.2
	75	83.4	69.9	49.1	23.0	90.9	71.4	40.7	14.3	95.0	
	100	85.5	70.4	48.9	22.6	91.1	81.3	53.5	50.6	6.8	97.2
ArcFace	5	99.2	97.1	89.2	60.0	76.1	94.8	77.6	37.4	1.8	98.5
	10	97.3	89.7	72.4	36.7	86.0	90.2	62.5	12.8	0.1	99.5
	25	91.2	80.7	60.3	29.1	88.3	87.4	46.2	12.9	0.6	99.4
	50	79.4	65.2	45.4	20.5	91.8	84.7	58.0	21.1	2.5	98.8
	75	72.2	57.5	37.3	16.8	93.4	71.4	40.7	12.9	0.8	99.4
	100	68.5	51.2	32.7	14.6	93.7	81.3	53.5	22.9	3.6	98.5
MagFace	5	99.2	97.7	91.3	65.1	72.3	99.8	98.2	89.7	59.3	80.7
	10	96.6	90.1	72.2	38.7	85.6	99.3	93.6	73.1	24.9	92.5
	25	82.8	70.1	49.1	22.1	90.8	94.2	81.4	42.9	3.4	98.0
	50	78.8	64.2	43.0	19.2	91.9	97.1	85.5	55.9	12.2	95.8
	75	73.4	57.0	37.3	17.5	92.5	97.3	85.6	56.6	10.7	96.1
	100	70.0	53.3	34.5	16.2	92.8	96.6	81.5	46.6	9.4	96.7
QMagFace	5	99.8	98.8	93.9	73.0	69.4	99.9	99.3	93.3	64.2	79.7
	10	96.5	89.5	72.3	39.7	84.4	99.3	93.1	71.6	20.1	93.7
	25	88.9	73.0	49.7	23.0	90.6	95.0	83.0	51.1	14.5	95.6
	50	79.6	62.8	41.9	19.9	91.3	98.5	86.7	57.8	14.2	95.3
	75	70.6	54.1	34.8	16.7	92.4	95.4	81.7	46.4	7.8	97.0
	100	73.8	56.1	37.0	17.3	92.7	96.3	81.4	54.0	14.8	95.2

Table 13. Conventional Open-Set Backdoor Performance on WebFace for the physical sunglasses trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the physical sunglasses trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).

		Error (FNMR) on Clean Data					Error (FNMR) on Backdoor Data				
	N	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC	10^{-4}	10^{-3}	10^{-2}	10^{-1}	AUC
FaceNet	5	98.6	95.2	84.8	55.4	77.2	99.5	97.8	90.5	59.0	81.4
	10	96.7	91.1	75.5	43.8	82.1	97.7	93.7	77.5	30.7	90.4
	25	92.5	83.3	62.7	32.9	85.9	98.9	95.0	78.9	27.5	91.7
	50	92.0	80.7	62.5	33.8	85.8	99.5	96.4	77.6	23.5	92.9
	75	94.4	84.4	65.8	37.3	85.0	99.7	96.9	80.2	27.4	91.6
	100	93.2	82.9	64.9	35.2	85.5	99.7	96.9	79.3	25.9	91.9
ArcFace	5	98.7	95.3	86.0	56.5	77.3	92.3	86.9	81.1	60.8	81.8
	10	98.4	94.9	83.0	51.9	79.3	92.4	85.2	54.3	5.4	97.3
	25	94.9	87.9	70.7	38.4	84.1	94.1	79.5	27.6	1.1	98.9
	50	88.6	76.5	56.0	27.7	88.8	76.4	43.3	8.9	0.4	99.6
	75	80.1	68.1	48.9	23.6	90.0	75.4	35.2	7.9	0.4	99.6
	100	79.0	65.6	45.9	22.2	90.2	64.6	34.8	5.9	0.1	99.8
MagFace	5	99.0	96.4	88.2	60.4	74.4	94.9	92.2	88.4	79.9	50.3
	10	96.6	89.9	72.7	38.7	84.5	95.9	93.6	91.6	87.9	48.3
	25	91.0	79.1	56.9	25.8	89.3	97.9	96.9	95.3	91.0	47.8
	50	94.7	85.6	66.1	32.3	87.3	99.4	98.6	95.4	82.5	55.4
	75	93.9	84.3	64.5	32.8	86.7	99.4	98.8	95.7	85.5	51.4
	100	95.7	87.6	70.6	38.6	84.1	99.7	99.1	95.9	82.4	56.2
QMagFace	5	99.3	97.2	88.5	60.0	75.4	99.3	96.6	86.6	80.7	43.4
	10	99.2	95.7	81.2	45.6	81.8	99.2	97.1	92.4	85.6	55.6
	25	96.6	88.0	67.9	32.2	87.4	99.0	97.7	95.9	89.8	48.1
	50	96.5	86.2	63.9	32.0	87.0	99.7	98.7	94.6	82.5	55.4
	75	95.1	86.0	63.2	31.0	87.6	99.7	98.8	95.8	83.5	54.3
	100	98.6	95.1	78.6	38.2	84.4	99.8	99.4	96.3	81.7	55.3

Table 14. Conventional Open-Set Backdoor Performance on WebFace for the red square trigger - In an open-set scenario, the conventional classification-based backdoors are evaluated based on FNMR[%]@FMR recognition error on the WebFace dataset for the red square trigger. A lower FNMR is better while the AUC should be higher. The Clean Data error refers to the evaluation using test data without poisoned images, while the Backdoor Data error is evaluated on data that consists of poisoned images. Generally, both on clean data and on backdoor data, the conventionally backdoored models show high error rates compared to the performance of the original model (see Tab. 5).