We are grateful to reviewers for the constructive comments, which help to improve the quality & clarity of the paper. Before addressing detailed comments, we summarize in Table 1 performances of the proposed methods under three ambiguity attack modes, $fake_i$ where $i = \{1, 2, 3\}$ depending on attackers' knowledge of the protection mechanism.

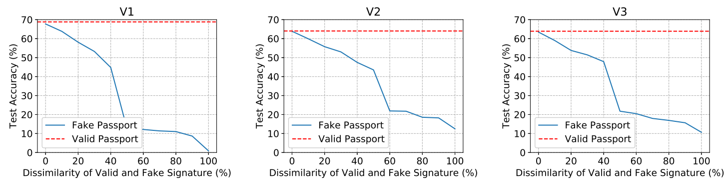

Figure 1: Test accuracy on CIFAR100 as suggested by R1 (i.e. try to create fake passport maximizing distance from P.)

Ambiguity attack modes	Attackers have access to	Ambiguous passport construction methods	Invertibility (see Def. 1.V)	Verification scheme V1	Verification scheme V2	Verification scheme V3
$fake_1$	W	Random passport P_r	$F(P_r)$ failed by big margin	Accuracy ↓ (68% → 1%) with fake passport.	Accuracy ↓ (65% → 1%) with fake passport.	Accuracy ↓ (65% → 1%) with fake passport.
$fake_2$	$W, \{D_r;D_t\}$	Reverse engineer passport P_e	$F(P_e)$ failed by moderate margin	Accuracy ↓ (68% → 30-45%) with fake passport.	Accuracy ↓ (65% → 20-30%) with fake passport.	Accuracy ↓ (65% → 20-30%) with fake passport.
$fake_3$	$W, \{D_r;D_t\}, \{P, S\}$	Reverse engineer passport $\{P_e;S_e\}$ by exploiting original passport P & sign string S	if $S_e = S$: $F(P_e)$ passed, with negligible margin* if $S_e \neq S$: $F(P_e)$ failed, by moderate to big margin	See Figure 1	See Figure 1	See Figure 1

Table 1: Performances(%) of V1, V2 and V3 schemes under three ambiguity attack modes, A.

where W are learned network weights; D_r, D_t are the training and testing datasets; $F()$ is the fidelity evaluation process, see Definition II in the main paper. * refer to S encodes ownership signature, which resolves the ambiguity.

In summary, when ambiguous passports are forged and used (*e.g. forge passport/watermark with the knowledge of verification method - R2 (see $fake_2$)*), Table 1 shows that all the corresponding network performances are deteriorated to various extent. The ambiguous attacks are therefore defeated according to the fidelity evaluation process, $F()$. We'd like to highlight that even under the most adversary condition i.e. $fake_3$ as suggested by R1, attackers are unable to change scale signs (which encode ownership information as detailed in supplementary Table 8) without compromising network performances. For example, with 10% and 50% of scale sign changes, the CIFAR100 classification accuracy drops about 5% and 50%, respectively. In case that the sign remain unchanged, network ownership can be easily verified by the pre-defined string of signs. Also, Table 1 shows that attackers are unable to exploit D_t to forge ambiguous passports (R2). We will include above results to the final draft.

Table 2 summarizes network complexity for various schemes. We believe it is the complexity and time cost during the inferencing stage that is to be minimized, since network inferences are to be performed frequently by end users. While extra costs at the training and verification stages, on the other hand, are not prohibitive since they are performed by network owners, with the motivation to protect network ownerships.

	V1	V2	V3
Training	- Passport layers added - Passports needed - 15-30% more training time	- Passport layers added - Passports needed - 100-125% more training time	- Passport layers added - Passports needed - Trigger set needed - 100-150% more training time
Inferencing	- Passport layers & passports needed - 10% more inferencing time	- Passport layers & passport NOT needed NO extra time incurred	- Passport layers & passport NOT needed NO extra time incurred
Verification	- NO separate verification needed	- Passport layers & passports needed	- Trigger set needed (black-box verification) - Passport layers & passports needed (white-box verification)

Table 2: Summary of network complexity for V1, V2 and V3 schemes.

R1: M_t is the network performance tested against D_t. The threshold ϵ_f is both datasets and network dependent, and has to be set empirically by network owners, to differentiate the genuine from fake passports. Theoretical analysis of either the threshold or its bounds might be a topic for future research.

R1: Evaluation of the cost of larger models. **Ans:** We tested a Resnet50 and its training time increases 10%, 182% and 191% respectively for V1.V2,V3 schemes. This increase is consistent with smaller models i.e. Alexnet and Resnet18.

R1: Other approaches to establishing ownership e.g. hosting models in trusted execution environments such as SGX enclaves? **Ans:** SGX enclaves is to ensure trusted execution of models without being tampered, while the proposed method is to protect the model from plagiarism (e.g. by a former staff who establish a new startup business with resources stolen from network owners).

R3: Fig 4-5 not clear. **Ans:** DNN performance is test accuracy. "valid" is test accuracy using valid passports and "orig" is baseline (unprotected model) test accuracy, we will correct "orig" to "baseline" instead.

R3: Performance (accuracy) suffer. **Ans:** Table 2 in the main paper shows the drop in test accuracy is no more than 1.5% for V1, V2, V3 compared with network without embedding any watermarks or passports.

R3: experiments uses V1, V2, V3. **Ans:** Results for all three schemes are presented in Table 2 from the main paper. We will also add detailed comparison as outlined by Table 1 & 2 above.

R3: We will revise the final submission as requested e.g. table headers/legends and other improvements. Thanks.