We are grateful to the referees for positive evaluations and useful suggestions for our work. Below we respond to the comments by each referee.

R1: It is possible to formulate our approximation result in terms of the number of weights. In fact, our construction of network is explicit and has number of weights = $N^2(L+1)$ where N is the width and L is the depth of the neural networks. This combines with Theorem 2.1 of the paper would lead to an estimate of number of weights needed to achieve certain approximation error. We will add a remark on this in the revised version. Thanks for the suggestion.

This paper only studies the approximation power of neural nets for generating distributions with an identical dimension. It is certainly very important and interesting to consider the dimension mismatch case, particularly the case where the dimension of the input distribution is much smaller than that of the output. However, constructing neural network based transport maps between generic distributions with unequal dimensions is highly non-trivial. Our result relies strongly on Brenier's theory of OT maps which is valid only in the equal-dimensional case. OT theory between unequal dimensions is rather challenging and characterizing the OT map in such case is far less understood. We will study this problem and report the results in future work.

Thanks very much for pointing out the relevant references. Let us briefly compare the results in those references with ours. Kong and Chaudhuri analyzed the expressive power of normalizing flow models under the L^1-norm of distributions and showed that the flow models have limited approximation capability in high dimensions. We show that feedforward DNNs can approximate a general class of distributions in high dimensions with respect to three IPMs. Lin and Jegelka studied the universal approximation of certain ResNets for multivariate functions; the approximation result there however was not quantitative and did not consider the universal approximation of ResNets for distributions. The work by Bailey and Telgarsky is closest to ours, but they only studied the approximation power of DNNs for expressing uniform and Gaussian distributions in the Wasserstein distance, whereas we proved quantitative approximation results for fairly general distributions under three IPMs. We will cite and comment on the references in the revised version.

R2: We listed the third point of the contribution in the paper to emphasize that the transport map we would like to approximate is highly structured, namely the maximum of finitely many affine functions. Thanks to the simple structure, the parameters of the ReLU DNN can be explicitly determined. We understand the referee's concern on this point as a contribution, and we will change this point as a remark on our results instead of a contribution in the revision.

We agree with the referee that Proposition 3.1 and Proposition 3.2 are known results. However, we state these results in a way that the dependencies of the convergence estimates on some key parameters (e.g. dimensionality) are more explicit than that in the literature. This is required for us to obtain explicit error estimates.

We understand that the referee feels that lines 248-274 about optimal transport might take too much space to be presented in an 8-page paper. We think however they are necessary for at least two reasons. First, since we are using optimal transport (with quadratic cost) to build the transport map between distributions, it is essential to recall the set-up of the optimal transport problem and the dual formulation, and also beneficial for readers without background knowledge on optimal transport. In addition, to describe the formulation of the optimal transport map in the semi-discrete case in Theorem 4.2, we need to introduce some notations such as the dual variable ψ and the power diagrams P_j. We also emphasize that a version of Theorem 4.2 for target measure defined on a compact convex domain was proved by Ref [20]; that result was however not enough for our purpose as the measures in our work are defined on the whole space $\mathbb{R}^d$. As a result, we come up with a different proof strategy based on purely the duality argument, instead of the geometric arguments used in Ref [20].

About Assumption K1 – Yes, the inequality should be strict. Thanks for catching this.

Thanks for pointing out the typos. We will correct them in the revised version. We will use $D(p, \pi)$ instead of $D(p, q)$ to avoid notation inconsistency. We require $k(x, y)$ to be twice differentiable since in the second condition of equation (2.5) we differentiate k with respect to both x and y. We will add the definition of $\mathcal{P}_2(\mathbb{R}^d)$ in the notation part of Section 1. We will use "absolute continuous with respect to Lebesgue" instead of "Lebesgue density" in the revision.

R3: We agree with the referee that the connection between optimal transport and generative models have been explored in the literature. To the best of our knowledge, however, this work is the first proof of a general and quantified universal approximation result of DNNs for distributions by using the optimal transport theory.

This work focuses mainly on the theoretical expressibility of DNNs for representing probability distributions. The transport map constructed in the paper is parameterized by the gradient of a DNN instead of a DNN itself and this gradient formulation originates essentially from the Brenier's theory for the optimal transport map. A discussion of the gradient formulation for practical training was included in the last paragraph of Section 2 and we plan to further expand that in the revision. Although the gradient parametrization is not as common as parametrizing the map itself, there have been an increasing number of works utilizing the gradients of neural networks for various practical leaning problems including GANs; see the discussion and the references mentioned at the beginning of page 6. Thanks for the comments.