We thank the reviewers for their positive view of our work and feedback. We address and clarify the items raised by the reviewers below. We will make sure to include the novel explanations and corrections in the camera ready version.

Novelty & Motivation: Transcription of piano performance video to music is a challenging task. State-of-the-art methods succeeded so far to predict coarsely onsets of notes, however, as we show, it is insufficient for realistically sounding music of the performance. With this work we thereby explore if it is *at all possible* to directly transcribe the music from video. Indeed, our work is the first full pipeline of transcribing (rather than generating similar) music with each component specifically designed to serve this purpose. Video2Roll is designed to capture detailed visual cues from video and transform them to binary prediction. We introduce components such as feature transform, feature refinement, and correlation learning which enhance Roll prediction. However, it is still a coarse binary prediction and does not directly correspond to pseudo-GT Midi (Figs. 5,6) critical for music synthesis. Our solution is to introduce a Roll to Midi context-conditional GAN, where given Roll, a Midi is generated to fool the discriminator. Even with the Midi, synthesis of music is not straightforward, since Midi is binary and missing expressive velocities. We thereby propose to use the same velocity to synthesize a mechanical audio, or PerfNet, a recent one-to-one deterministic transcription from Midi to Spectrogram. While the methods that we compose into the first two pipeline components, such as multi-scale features, self-attention mechanism, conditional GAN, Unet, have been generically introduced, they have not been systematically combined to transcribe music and as far as we know in any video-audio system. Our work thus identifies such necessary components (out of many examined) to experimentally succeed in music transcription.

Possible Applications: An important guideline for our pipeline was interpretability and modularity of implementation such that it would be possible to incorporate it in various video-music applications with piano. An immediate type of application would be on-the-fly music transform (e.g., we show a timbre transform in the paper). Adding a camera on top of a general piano keyboard (no need for electronic) could generate various timbres and can be possibly implemented in real-time application. An extension of such a real-time application would be a virtual piano, where in a virtual-reality environment, without need for mechanical instrument at all, the pipeline could produce a full virtual piano experience. In addition, we foresee applications that analyze video-audio streams in postprocessing manner. For example, as R2 suggests, with mounting a camera on top of the piano it could be possible to isolate the piano transcription from a multi-instrument performance, without affecting the performance, or as R1 suggests, our pipeline may be combined with current audio-only piano transcription methods. Indeed, as we discussed in the paper, the additional visual cues detected and processed by the pipeline could be matched for audio-visual synchrony and enhance the output.

Generality: In contrast to many previous works designed or tested in a specific lab setting, we aim to use general top-view Youtube videos not recorded for the purpose of our method with no specifics on the recordings settings. Indeed, the instrument and camera setup are *not required to be the same* for the recordings that we used. The performer played music on different pianos and aspect ratio of the keyboard under the camera was variable. During the pre-processing step, we implement elimination of biases such as color, piano shapes, by setting all frames to gray scale, crop of keyboard region, and transformation to common frame size (100×900). One unavoidable bias is of the performer/s hands. However, since Video2Roll network is designed to focus on generic visual cues and not on the hand specifics, we expect that the importance of hands would not be significant when additional performers are in the dataset. Notably, we addressed other aspects of generalization such as variation in music style. We made sure that when training is limited to a single composer (e.g. Bach) our pipeline is tested on a variety of music styles and would transcribe music with similar quality as of composer used for training (e.g., see supplementary video *NeurIPS2020_2811_sup_video.mp4*). Our experiments suggest that additional videos would not necessarily improve precision. We foresee significant possible enhancement if the dataset is curated to have balances in terms of variations of keys to be detected.

Response to R1: Please see clarifications regarding motivation, possible applications and generalizations above. The Onsets-and-Frames(OF) framework has high average frame-level precision (88.5% without velocity prediction) on MAPS dataset. OF allows us to use videos from YouTube. We only use the binary representation, since OF has a low precision on expressive velocity prediction (35.52%). The imperfect pseudo GT may indeed impair evaluation when done on Midi only and not tightly related to the transcribed music. We thereby include additional audio evaluation protocols (SoundHound and Human evaluation) (see line 231 in paper and supp. materials). As an initial work, we do not consider pedaling since it would be necessary to include an annotated video-audio dataset to explore this feature.

Response to R2: Please see clarifications regarding applications and generalizations above. We will elaborate further on the audio synthesis portion which deals with the problem that pseudo Midi (binary & missing velocities) synthesis to music is not straightforward. We propose to use the same velocity to synthesize a mechanical audio or to learn the expressive velocities implicitly via NN-based synthesizer (PerfNet).

Response to R3: We added the citations to Kidron et. al., 05' and Barzelay et. al., 07' on classical methods and plan to discuss them in the Related Work section of the camera ready version. We will also add details on system choice and fix the errors mentioned in the feedback.

Response to R4: Estimating expressive velocities requires to have a Midi GT which specifies them, however the Onsets-and-Frames(OF) that we use as GT generates velocity prediction with 35.52% precision. We therefore propose instead to learn it implicitly via a NN-based synthesizer (PerfNet). Similarly for rhythm and pedaling, these would require high enough quality annotated GT such that they would add to enhanced audio output beyond PerfNet.